DIGITAL CURATION
PRACTICE, PROMISE & PROSPECTS

Edited by:
Helen R. Tibbo
Carolyn Hank
Christopher A. Lee
Rachael Clemens

April 1-3, 2009
University of North Carolina at Chapel Hill, NC USA
Proceedings of DigCCurr2009
Digital Curation: Practice, Promise and Prospects

This volume contains contributions to DigCCurr2009, held April 1-3, 2009 at the University of North Carolina at Chapel Hill.

This work is licensed under the Creative Commons Attribution-Noncommercial-No Derivative Works 3.0 Unported license. You are free to share this work (copy, distribute and transmit) under the following conditions: attribution, non-commercial, and no derivative works. To view a copy of this license, visit http://creativecommons.org/licenses/by-nc-nd/3.0/.

ISBN 978-0-578-02215-4

DigCCurr2009, April 1-3, 2009, Chapel Hill, NC, USA

School of Information and Library Science
University of North Carolina at Chapel Hill
CB #3360, 100 Manning Hall
Chapel Hill, NC 27599-3360
Phone: 919.962.8366
Email: info@ils.unc.edu
Preface

It is our pleasure to introduce you to the Proceedings of DigCCurr2009: Digital Curation Practice, Promise and Prospects.

DigCCurr2009 was held on April 1-3, 2009 in Chapel Hill, North Carolina as part of the Preserving Access to Our Digital Future: Building an International Digital Curation Curriculum (DigCCurr) project. DigCCurr is a three-year (2006-2009), Institute of Museum and Library Services (IMLS)-funded project to develop a graduate-level curricular framework, course modules, and experiential components to prepare students for digital curation in various environments. DigCCurr initiatives in support of this goal are informed by representatives from the project’s collaborating institutions as well as an Advisory Board of experts from Australia, Canada, Italy, the Netherlands, New Zealand, the United Kingdom and the United States.

The first symposium, DigCCurr2007: An International Symposium in Digital Curation, was held April 18-20, 2007, attracting nearly 300 attendees from ten countries. Participants explored the definition of digital curation and what skills are necessary for digital curation professionals working in libraries, archives, museums, data centers, and other data-intensive organizations.

DigCCurr2009 continues the same general theme as the symposium two years earlier, focusing on current practice and research surrounding digital curation with a look toward the future, and trends in preparing digital curation professionals.

The call for participation encouraged submissions on a wide range of topics, including but not limited to digital curation synergies and collaboration; teaching and training at the international level; digital curation in relation to archives and museums; current digital curation practices; pressing digital curation needs; and infrastructures in support of digital curation.

Contributions to DigCCurr2009 take the form of short and long papers, posters and panels. Potential contributions were submitted for peer review by a rich and diverse panel of international experts. Reviewers evaluated the submissions based on clarity and organization of presentation and writing; originality, creativity and potential for new contributions to the field; and engagement (topics addressed would be appropriate for and engaging to the diverse audience of DigCCurr2009 participants). Authors of accepted submissions had the opportunity to revise their documents before final production of the proceedings. We would like to extend our gratitude to all of the reviewers. Their efforts have contributed substantially to the quality of the conference and proceedings.

The DigCCurr2009 program was also greatly enhanced by several invited papers and panels by a distinguished body of international experts. These invited papers and panels were not subject to the same peer review process as the other contributions, and you will note that their titles begin with the word “invited” to reflect this.

We introduced a special panel format for the DigCCurr2009 called “Tools and Demos.” In these sessions, several individuals had the opportunity to speak for just a few minutes about a specific digital curation tool, and then the remainder of the time was available for audience members to move around the room and experience the tools more directly. This was a great opportunity to expose DigCCurr2009 attendees to many of the leading-edge resources to support digital curation work.
The production of this proceedings volume is the result of efforts by numerous individuals. We would particularly like to thank Sarah Carrier, Rachael Clemens, Lori Eakin, and Lisa Gregory for their editing and producing the text that you see here, and Songphan Choemprayong for cover design.

The DigCCurr2009 program included eighteen panels, twenty seven papers, thirteen demos, and nine posters. The keynote address, “Building the Universal Library: The Promise and Challenges of HathiTrust,” was delivered by John Wilkin, Associate University Librarian for Library Information Technology and Technical and Access Services, University of Michigan Library. In total, there were more than a hundred presenters, coming from nine different countries.

We hope that you gain from reading these proceedings and that participants in DigCCurr2009 all benefited from the event. It has been a true pleasure to work with and draw expertise from such a great set of people.

Christopher (Cal) Lee
Helen R. Tibbo
Conference Chairs

Carolyn Hank
Program Chair

School of Information and Library Science
University of North Carolina, Chapel Hill
# TABLE OF CONTENTS

**DigCCurr2009 Organization** ........................................................................................................ x  

**DigCCurr Project Organization** ...................................................................................................... xii  

**Concurrent Session 1:**  

A. **Common Workflows: Health and Social Science Data Curation Collaborations** ............................... 1  
   Granda, P., Thomas, D., Grasso, C., & Teixeira, C.  

B. **Comparing Curricula for Digital Library and Digital Curation Education** ........................................ 2  
   Pomerantz, J., Oh, S., Wildemuth, B.M., Hank, C., Tibbo, H.R., Fox, E.A., & Yang, S.  

C. **Distributed Custodial Frameworks for Archival Preservation** ................................................................. 4  
   Marciano, R., Wojcik, C., Wilczek, E., Conrad, M., & Tibbo, H.R.  

D. **Funders’ Perspectives** ....................................................................................................................... 6  
   Wurl, J., Ray, J., Grindley, N., & Williams, K.  

**Concurrent Session 2:**  

A. **Education for Digital Curation:**  

   - **Thinking Like a Digital Curator: Creating Internships in the Cognitive Apprenticeship Model** ........................................................................................................ 7  
     Yakel, E., Conway, P., & Krause, M.G.  
   
   - **Qualification & Education in Digital Curation: the nestor Experience in Germany** .......................... 12  
     Neuroth, H., Osswald, A., & Strathmann, S.  
   
   - **Educating Archivists about Copyright: How Can We Do It Better?** .................................................... 19  
     Dryden, J.  
   
   - **Building Australia’s eResearch Capability: The Challenge of Data Management** ............................ 24  
     Burton, A. & Henty, M.  

B. **Recordkeeping and Government Information Curation:**  

   - **The Survival of Records (and Records Management) in the Twenty-First Century** ........................... 26  
     Thibodeau, K.  
   
   - **Pathways to Preservation: Digital Curation Strategies in North Carolina State Government** ............. 31  
     Eubank, K., Ricker, J., & Rudersdorf, A.  
   
   - **MoReq2: a European Contribution to the Preservation of Electronic Records** ............................... 37  
     Fresko, M.
C. Digital Curation of Humanistic, Multimedia Materials: Lessons Learned and Future Directions .......................................................... 42
Winget, M., Frick, C., McDonough, J., Rennear, A., & Lowood, H.

D. Digital Curation Tools and Demos I .......................................................... 44
Rauber, A., van der Hoeven, J., van Diessen, R.J., Pearce-Moses, R., Bowden, H., & Pomerantz, J.
  • Creating a Preservation Plan Using the Preservation Planning Tool Plato .................................................. 45
    Kulovits, H., Becker, C., Kraxner, M., & Rauber, A.
  • Hoppla - Digital Preservation Support for Small Institutions .................................................. 47
    Strodl, S., Motlik, F., & Rauber, A.

Concurrent Session 3:

A. Metadata:
  • Effective Access to Digital Assets: An XML-based EAD Search System ............................................... 49
    Zhang, J., Fachry, K.N., & Kamps, J.
  • Integrating Metadata into the NARA Transcontinental Persistent Archive Prototype via the OAI-PMH .......................................................... 57
    Ward, J., de Torcy, A., Mantooth, J., Chua, M., & Crabtree, J.
  • Creating Metadata for a Digital Database: A Case Study .......................................................... 60
    She, J. & Chace, M.B.

B. Curation of Scientific Datasets: Trends, Current Initiatives, and Solutions ........................................... 68

C. Technology Learning for Digital Curators ................................................................................. 70
Botticelli, P., Bradley, J., & Fulton, B.

D. Digital Curation Tools and Demos II .................................................................................. 72
Pearson, D., Shah, C., Moore, R., Ingram, G.B., McHugh, A., & Hofman, H.
  • Prometheus: Managing the Ingest of Media Carriers .......................................................... 73
    del Pozo, N., Elford, D., & Pearson, D.
  • Mediapedia: Managing the Identification of Media Carriers .................................................. 76
    del Pozo, N., Elford, D., & Pearson, D.
  • ContextMiner: Collect Different .................................................................................. 79
    Shah, C.
  • Teaching with CONTENTdm in the Digital Curation Curriculum ........................................... 81
    Ingram, G.B.
Concurrent Session 4:

A. Digital Curation and Preservation Training and Education: A Panel to Consider Options and Intersections ................................................................. 83
   McGovern, N.Y., Tibbo, H.R., Cragin, M.H., Davidson, J., & Hofman, H.

B. Snapshot of Digital Preservation in Federal Libraries ............................................. 85
   Keller, D., Murphy, P, Dunn, C., Huffine, R., & Crawford, C.

C. Digital Curation Vignettes: Personal, Academic, and Organizational Digital Information .............................................................................................................................. 87
   Beaudoin, J.E., Esteva, M., & Japzon, A.

D. Cooperative Approaches to Digital Preservation .................................................. 90
   Halbert, M., Walters, T., Trehub, A., Pearce-Moses, R., & Crabtree, J.

Concurrent Session 5:

A. Extending the Data Curation Curriculum to Practicing LIS Professionals .................. 92
   Cragin, M.H., Smith, L.C., Palmer, C.L., & Heidorn, P.B.

B. Archives in the Wild:

   • Everyone is a Curator: Human-Assisted Preservation for ORE Aggregations ................. 94
     McCown, F., Nelson, M.L., & Van de Sompel, H.

   • Digital Curation and the Citizen Archivist .................................................................... 102
     Cox, R.J.

   • Use of Computer Forensics in the Digital Curation of Removable Media ......................... 110
     Glisson, W.B.

C. Moving Web Archiving into the Classroom ................................................................ 112
   Bragg, M., Fox, E.A., Hedstrom, M., & Lee, C.A.

D. Gaps and Persistent Challenges ................................................................................. 114
   Lynch, C., Sawyer, D., Ashley, K., & van Diessen, R.J.

Concurrent Session 6:

A. Personal Digital Archiving ....................................................................................... 115
   John, J.L., Marshall, C.C., Pearson, D., & Rauber, A.

B. Digital Curation Policy Issues ................................................................................... 117
   Smith, C.A., Agnew, G., Cragin, M.H., Dryden, J., & Eschenfelder, K.R.

C. Scientific and Humanities Data Curation:

   • “Able To Develop Much Larger and More Ambitious Projects”: An Exploration of Digital Projects Teams ................................................................. 119
     Siemens, L., Duff, W., Cunningham, R., & Warwick, C.
• Documentation Evaluation Model for Social Science Data: An Empirical Test ........................................ 125
  Niu, J., & Hedstrom, M.

• Data Access and Long-Term Data & Knowledge Preservation for Earth Science: An Overview on Some ESA Initiatives ............................................................... 130
  Albani, S., Beruti, V., Fusco, L., & Giaretta, D.

D. Digital Curation Research:

• Speech Acts and Electronic Records .......................................................................................... 135
  Underwood, W.E.

• Reconstructing the Digital Past: A Case Study of the Reconstruction of the Lost Pittsburgh Project ........................................................................................................ 143
  Callery, B.G.

• Identifying and Implementing Modular Repository Services .................................................. 145
  Johnston, L.

Concurrent Session 7:

A. Digital Curation for Cultural Institutions

• Web Access for the Museum of Anthropology’s Collection ...................................................... 149
  Whittington, S.L., Bryner, K.E., Hancock, B.H., & Vidrine, M.R.

• Communicating Archives of Cultural Institutions: Venice as a Case Study ...................................... 154
  Niero, M. & Urbani, C.

• The Russian Doll Effect: A Case Study in Digital Artifact Recontextualization ............................ 156
  Bradley, J.

• Capturing a Plurality of Perspectives: A Framework for Developing Culturally Sensitive Curriculum and Digital Repositories ...................................................... 158
  White, K.L. & Abbas, J.

B. Skills for Significant Properties: Debating Pragmatics and Philosophy in an Area of Digital Curation ........................................................................................................ 161
  Grace, S., Anderson, S., & Lee, C.A.

C. Change Management ...................................................................................................................... 163
  Eaton, F., Chapman, S., & Crabtree, J.

D. Digital Curation Tools & Strategies

• An Implementation of the Audit Control Environment (ACE) to Support the Long Term Integrity of Digital Archives ................................................................. 164
  Smorul, M., Song, S., & JaJa, J.
• A Digital Library Service for the Small .............................................................. 170
  Angelis, S., Constantopoulos, P., Gavrilis, D., & Papatheodorou, C.

• Lessons Learned from the DISC-UK DataShare and Data Audit Framework Implementation Projects ............................................................ 175
  Rice, R.

• Preservation Workflows, Strategies and Infrastructure .................................... 177
  Giaretta, D.

Poster Session:

A Residential Data Curation Internship: Opportunities and Challenges .................... 185
Gordon, K.M.

Preserving Electronic Mailing Lists: The H-Net Archive ........................................ 187
Schmidt, L.M.

Federal Libraries Digital Preservation Census .................................................... 189
Keller, D. & Harrison, A.

Extending an LIS Data Curation Curriculum to Include Humanities Data ................ 191

Data Management and Curation of Research Data in Academic Scientific Research Environments ................................................................. 194
Hayes, B.E., Harroun, J.L., & Temple, B.

Contextual Information from Blogs in Video Digital Curation ............................ 196
Capra, R.G., Clemens, R., Lee, C.A., & Sheble, L.

What Should We Teach about METS in a Digital Preservation Course? ................ 199
Waters, J. & Allen, R.B.

Eakin, L.

Getting the Tar Off Our Heels: Moving Forward with Archiving University of North Carolina at Chapel Hill Websites ....................................................... 204
Gregory, L.

Author Index ........................................................................................................ 207
DIGCCURR 2009 ORGANIZATION

Planning Committee

Helen R. Tibbo, University of North Carolina, Chapel Hill
Christopher A. Lee, University of North Carolina, Chapel Hill
Carolyn Hank, University of North Carolina, Chapel Hill
Rachael Clemens, University of North Carolina, Chapel Hill
Wendy Duff, University of Toronto
Maria Guercio, University of Urbino
Seamus Ross, University of Toronto
Kenneth Thibodeau, National Archives and Records Administration
Elizabeth Yakel, University of Michigan

Reviewers

Ian Anderson, University of Glasgow
Deborah Barreau, University of North Carolina, Chapel Hill
Priscilla Caplan, Florida Center for Library Automation
Stephen Chapman, Harvard University Library
Sayeed Choudhury, Johns Hopkins University
Rachael Clemens, University of North Carolina, Chapel Hill
Paul Conway, University of Michigan
Richard Cox, University of Pittsburgh
Adrian Cunningham, National Archives of Australia
Amber Cushing, University of North Carolina, Chapel Hill
Robin Dale, University of California, Santa Cruz
Raymond J. van Diessen, IBM Business Consulting Services
Wendy Duff, University of Toronto
Edward Fox, Virginia Tech
Neil Grindley, Joint Information Systems Committee
Maria Guercio, University of Urbino
Carolyn Hank, University of North Carolina, Chapel Hill
Margaret Hedstrom, University of Michigan
Bradley Hemminger, University of North Carolina, Chapel Hill
Hans Hofman, Nationaal Archief Netherlands
Christopher A. Lee, University of North Carolina, Chapel Hill
William LeFurgy, Library of Congress
Gary Marchionini, University of North Carolina, Chapel Hill
Richard Marciano, University of North Carolina, Chapel Hill
Jerome McDonough, University of Illinois, Urbana-Champaign
John McMullen, University of Illinois, Urbana-Champaign
Andrew McHugh, University of Glasgow
Michael Moss, University of Glasgow
Javed Mostafa, University of North Carolina, Chapel Hill
Jeffrey Pomerantz, University of North Carolina, Chapel Hill
Soo Young Rieh, University of Michigan
Donald Sawyer, VIE Inc.
Helen R. Tibbo, University of North Carolina, Chapel Hill
Elizabeth Yakel, University of Michigan
Proceedings Support
Sarah Carrier, University of North Carolina, Chapel Hill
Rachael Clemens, University of North Carolina, Chapel Hill
Lori Eakin, University of North Carolina, Chapel Hill
Lisa Gregory, University of North Carolina, Chapel Hill
Carolyn Hank, University of North Carolina, Chapel Hill

Graphic Design
Songphan Choemprayong, University of North Carolina, Chapel Hill

Sponsors
UNC
SCHOOL OF INFORMATION AND LIBRARY SCIENCE

UNC
UNIVERSITY LIBRARY

INSTITUTE of Museum and Library SERVICES

National Archives and Records Administration (NARA)

CNI
Coalition for Networked Information
DIGCCURR PROJECT ORGANIZATION

DigCCurr Advisory Board
Stephen Chapman, Harvard University Library
Adrian Cunningham, National Library of Australia
Robin Dale, University of California, Santa Cruz
Raymond van Diessen, IBM Business Consulting Services
Wendy Duff, University of Toronto
Philip Eppard, State University of New York, Albany
Anne Gilliland, University of California, Los Angeles
Maria Guercio, University of Urbino
Hans Hofman, Nationaal Archief Netherlands
Anne Kenny, Cornell University Library
Steven Knight, National Library of New Zealand
Clifford Lynch, Coalition for Networked Information
Richard Marciano, University of North Carolina, Chapel Hill
Seamus Ross, University of Toronto
Donald Sawyer, VIE Inc.
Kenneth Thibodeau, National Archives and Records Administration
Elizabeth Yakel, University of Michigan

DigCCurr Project Team, UNC-Chapel Hill
Helen R. Tibbo, PI
Christopher A. Lee, Co-PI
Gary Marchionini, Co-Investigator
Jeffrey Pomerantz, Co-Investigator
Carolyn Hank, Project Manager

Carolina Digital Curation Masters’ Fellows
John Blythe
Alexandra Chassanoff
T. Mike Childs
Lisa Gregory
Samantha Guss
Jennifer Mantooth
Katherine-Rose Repp
Mark Swails
Tara Wink

Local Partners, UNC-Chapel Hill
ibiblio.org
Information Technology Services’ Teaching and Learning Division
The Odum Institute for Research in Social Science
University Libraries
Common Workflows: Health and Social Science Data Curation Collaborations

Peter Granda, David Thomas  
ICPSR  
University of Michigan  
Ann Arbor, MI 48109 USA

c PeterG@umich.edu  
davethom@isr.umich.edu

Chris Grasso  
The Fenway Institute  
Fenway Community Health  
Boston, MA 02115 USA

Grasso@fenwayhealth.org

Cindy Teixeira  
The Roper Center  
University of Connecticut  
Storrs, CT 06269 USA

cindy@ropercenter.uconn.edu

ABSTRACT
This panel session will describe collaborations between different organizations to establish common technologies and procedures to process and preserve quantitative data in health and the social sciences for dissemination to the research community.

In normal practice data producers and archives work independently to produce public-use data and documentation files for their users. They often use similar procedures for creating metadata, cleaning and checking data values, and making the final products accessible to others. Our panel will discuss the ramifications that such an environment has had on the availability and quality of research data.

It is now the case that the health and social science data communities have developed some common standards for much of the work they do in the areas of data preservation, processing, and dissemination. Our panel will describe some of these standards and focus on two specific collaborations involving data in the areas of health and public opinion research. We will discuss the ways that these collaborations produce common workflows that leverage existing expertise, save staff resources, and generate enhanced data resources for students, teachers, and researchers.

ICPSR is collaborating with the Fenway Institute in Boston, MA to disseminate data to the research community that deals specifically with Lesbian, Gay, Bisexual, and Transgender populations. Unlike the normal relationship ICPSR has with data producers where ICPSR staff does the actual processing work, the staff of the Fenway Institute remotely accesses processing tools located at ICPSR and utilizes ICPSR methodology to archive their own data with ICPSR.

The collaboration between ICPSR and the Roper Center for Public Opinion Research at the University of Connecticut is a joint venture to process polling data traditionally processed and disseminated by each organization. Initially focusing on the CBS News and ABC News poll data collections, the collaboration leverages the automated data processing systems of the ICPSR and the commercial polling data expertise and focus of the Roper Center. The goals of this collaborative effort are: 1) a single stream processing of studies, eliminating duplicate effort by both organizations, 2) to produce the same files and collection information for each study, 3) to create standardized metadata for each study providing references and links between the same files within each archive, and finally, 4) to produce joint citation information.

Chris Grasso of the Fenway Institute will discuss the standards and concerns of archiving medically related data to create a comprehensive public access research data library, accessible on the Internet, for researchers around the world. Disclosure, in of itself is a very serious matter, but may have dire consequences in communities that may be subject to biases based on the information disclosed.

Peter Granda and David Thomas will provide an overview of how ICPSR prepares data for dissemination for the purposes of secondary analysis. Along with Cynthia Teixeira, they will discuss the importance of opinion polls, how they acquire such polls, and the benefits of collaborating on preparing public opinion polls.

Categories and Subject Descriptors

Keywords
Collaboration, preservation, dissemination.
Comparing Curricula for Digital Library and Digital Curation Education

Jeffrey Pomerantz, Sanghee Oh, Barbara M. Wildemuth, Carolyn Hank, Helen Tibbo
School of Information and Library Science
University of North Carolina at Chapel Hill
Chapel Hill, NC 27599 USA
{pomerantz, shoh, wildem, hcarolyn, tibbo}@unc.edu

Edward A. Fox, Seungwon Yang
Department of Computer Science
Virginia Tech
Blacksburg, VA 24061 USA
{fox, seungwon}@vt.edu

ABSTRACT
Two related curriculum development projects are currently underway, one concerning digital libraries and one concerning digital curation. This paper explores the convergence and divergence of these two federally-funded projects' approaches to curriculum development.

Categories and Subject Descriptors
K.3.2 [Computers and Education]: Computer and Information Science Education – computer science education, curriculum, information systems education.

Keywords
Digital libraries, digital librarianship, digital curation, computer science, library and information science, education, curriculum

1. INTRODUCTION
The two closely-related fields of digital libraries and digital curation have developed rapidly over the past few decades. At this point, both are in need of curriculum development efforts, to support the education of practitioners in each area.

Hundreds of millions of dollars have been invested in digital library (DL) research since the early 1990s, including research on how DLs can aid education, but investment to support teaching and learning about DLs has not kept pace. The Computing Curriculum 2001 (CC2001) defines curricula for Computer Science [1], and includes DLs as one of 14 knowledge modules under Information Management, emphasizing the importance of the topic in CS education.

Six years after the “It’s About Time” report [4], the need for education and training of digital curation professionals is more pressing than ever. Several disciplines and professions have developed de facto practices and expertise in aspects of digital curation, but without guiding principles or an overarching vision of data preservation and reuse. Professional education for digital curation has generally involved on-the-job training and experimentation, sometimes supplemented by workshops.

In response to the need for graduate-level education to prepare future professionals for DL and curation work, two projects are currently underway: one to develop a DL curriculum and one to develop a curation curriculum. Given their common concern with the management of online repositories of digital objects, these two projects have been collaborating since their inception. Indeed, the distinction between a library and an archive is not always clear, and when made available online, the distinction is even less clear. It has therefore been a concern for both projects to articulate how the scope of its work is similar to or differs from that of the other project. This paper briefly explores how these two projects converge and diverge, both in their underlying principles and in their approaches to curriculum development.

2. BACKGROUND
2.1 Digital Library Curriculum Project
The University of North Carolina at Chapel Hill School of Information and Library Science (SILS) and the Virginia Tech Department of Computer Science are currently developing a curriculum for teaching DL topics1. A central goal of this project is to develop curricular materials that may be used in both Information and Library Science (ILS) and Computer Science (CS) programs. Based in part on analyses of DL course syllabi in ILS and CS programs, and analyses of published papers in DL-related journals and conferences [5, 6], a framework has been developed for these curricular materials that outlines a recommended sequence for teaching these modules in a two-semester program. The modules that have been developed are the equivalent of a lesson plan for a 1.5- to 3-hour class session. Several curriculum modules have been developed and are being field tested by instructors in DL courses; more are currently being developed and evaluated. The framework of the proposed

---

1 Supported by the National Science Foundation under Grant Nos. IIS-0535057 (VT) and IIS-0535060 (UNC-CH).
While both projects address all stages of the digital object lifecycle, each concentrates on specific parts. The DL Curriculum Project framework focuses primarily on two stages, represented in the DCC Model as (1) the sequential actions from creation to transformation of digital objects, and (2) the lifecycle-long action of assignment of description and representation information. The DL Curriculum Project also focuses on characteristics of the digital object itself. The DigCCurr Project high-level categorization focuses primarily on preservation planning, and curation and preservation activities, as well as migration.

The DL Curriculum Project framework outlines 10 “core” topics in DL education, with 47 individual modules within these umbrella topics. The DigCCurr matrix of knowledge and competencies outlines 23 individual units within six dimensions. While no mapping between frameworks has yet been developed, there are clear intersections and equally clear differences. As both projects continue to develop curricular materials, it will become apparent where modules and units overlap.

4. CONCLUSION

The DL Curriculum Project and the DigCCurr Project address similar problems of developing curricula to prepare professionals to manage repositories of digital objects. The overlap between the scopes of these projects makes collaboration essential. The differences in the focus and approaches of these projects, however, make it clear that, in order to develop a curriculum that addresses the full lifecycle of digital objects, a diversity of perspectives is necessary. DLs and digital curation are two such perspectives; others, for example museum informatics, may also inform this curriculum development effort in the long term. These curricular concentrations have the potential to significantly shape the future of ILS and CS programs. The authors will continue to identify areas where one project will support the goals of the other project, and of ILS and CS education.

5. REFERENCES


2 Supported by IMLS Grant No. RE-05-06-0044.
Distributed Custodial Frameworks for Archival Preservation

Richard Marciano
School of Info. & Library Science
Univ. of North Carolina at Chapel Hill
Chapel Hill, NC 27599 USA
richard_marciano@unc.edu

Mark Conrad
Electronic Records Archive
National Archives and Records Administration
Rocket Center, WV 26726 USA
mark.conrad@nara.gov

Caryn Wojcik
Records Management Services
State of Michigan
Lansing, MI 48909 USA
wojcikc@michigan.gov

Helen R. Tibbo
School of Info. & Library Science
Univ. of North Carolina at Chapel Hill
Chapel Hill, NC 27599 USA
tibbo@email.unc.edu

Eliot Wilczek
Digital Collections and Archives
Tufts University
Medford, MA 02155 USA
eliot.wilczek@tufts.edu

ABSTRACT

Building end-to-end digital repositories is a task few institutions can currently accomplish. The Distributed Custodial Archival Preservation Environments (DCAPE) project addresses this problem by offering a preservation service that includes a trusted digital repository infrastructure assembled from state-of-the-art rule-based data management systems, commodity storage systems, and sustainable preservation services. Presenters discuss case studies of preservation environments they are jointly building. These include TPAP (Transcontinental Persistent Archives Prototype), DCAPE, Fedora / iRODS integration, and iRODS (integrated Rule-Oriented Data Systems) prototypes and represent distributed custodial frameworks from the federal, state, university, and cyberinfrastructure perspectives.

Categories and Subject Descriptors
D.3.3 [Information Storage and Retrieval]: Systems and Software – distributed systems, information networks.

Keywords
Digital preservation, archives, data grids, policy-oriented data management.

1. INTRODUCTION

This paper addresses the main Annual Meeting theme of sustainability. The goal is to re-imagine the possibilities of distributed custody in archives and the potential collaborative and business models that result and can help address the challenge of long-term digital preservation. The perspectives of distributed custody from the federal (Mark Conrad), state (Caryn Wojcik), university (Eliot Wilczek), and cyberinfrastructure (Richard Marciano) viewpoints are represented, thus providing a unified set of themes. Distributed data management systems that promote automation and customization are discussed and the overall proceedings are moderated by Helen Tibbo.

2. PRESERVATION CASE STUDIES

The case studies represented by the different speakers include:

iRODS [1]: The integrated Rule-Oriented Data System is adaptive middleware that provides a flexible, extensible and customizable data grid architecture. It supports extensibility and customizability by encoding operations into sequences of micro-services. It is a data grid software system developed by the Data Intensive Cyber Environments (DICE) [2] group (developers of the SRB, the Storage Resource Broker), and collaborators. The iRODS system is based on expertise gained through nearly a decade of applying the SRB technology in support of Data Grids, Digital Libraries, Persistent Archives, and Real-time Data Systems. The most important change is the introduction of management policies (sets of assertions these communities make about their digital collections) which are implemented as machine-actionable rules and state information. At the iRODS core, a Rule Engine interprets the rules to decide how the system is to respond to various requests and conditions. Finally, iRODS is open source under a BSD license.

TPAP [3]: The Transcontinental Persistent Archives Prototype (TPAP) is a research testbed developed in partnership between the National Archives and Records Administration (NARA), the Data Intensive Cyber Environments (DICE) Center at the University of North Carolina, Chapel Hill (UNC-CH) and at the University of California, San Diego (UCSD), and the University of Maryland Institute for Advanced Computer Studies (UMIACS). Its aim is to research and develop theories, concepts, and technologies for distributed long-term digital data preservation. TPAP is a key component of the infrastructure in the ERA Research Laboratories at NARA and is used as a prototype to inform research for the
The goal is to build a preservation service that meets the needs of archival repositories for trusted archival preservation so they do not have to build their own in-house electronic records archives. The preservation service will allow the archival repositories to distribute physical custody, while retaining legal custody. The preservation environment employed by the service will build upon the technologies developed at UNC at the Renaissance Computing Institute (RENCI). The preservation environment includes a trusted digital repository infrastructure that is assembled from state-of-the-art rule-based data management systems, commodity storage systems, and sustainable preservation services. The software infrastructure automates many of the administrative tasks associated with management of archival repositories and validation of trustworthiness. RENCI and faculty at SILS/UNC are working with expert archivist partners from state archives, university archives and cultural archives to demonstrate the viability of the distributed custodial preservation approach as a production service. We have assembled a partnership of 28 people across 9 institutions and 4 staff at UNC, for a total of 32 participants. The institutions are:

- **State Archives**: California, Kansas, Michigan, Kentucky, North Carolina, New York
- **University Archives**: Tufts University
- **Cultural Entity**: Getty Research Institute.
- **Cyberinfrastructure Partners**: West Virginia University (WVU), Renaissance Computing Institute (RENCI), School of Information and Library Science (SILS) at UNC, and Carleton University (Canada)

DCAPE officially started in December 2008, and preliminary findings and new directions will be discussed.

**Fedora and iRODS integration** [5]: Integration efforts are under way to connect Fedora and iRODS. In this approach, iRODS is used for storing Fedora objects, and Fedora is used for metadata/object management. iRODS provides a true distributed storage environment for Fedora. It can manage Fedora objects and distribute them geographically.

This integration project is being tested through the Carolina Digital Repository (CDR) initiative at UNC Chapel Hill Libraries.

### 3. Acknowledgments

The research results in this paper were partly funded by the NSF Office of Cyberinfrastructure OCI-0848296 grant, “NARA Transcontinental Persistent Archives Prototype”, (2008-2012). The iRODS technology development is funded by NSF SDCI 0721400, “SDCI Data Improvement: Data Grids for Community Driven Applications” (2007-2010).


### REFERENCES


Invited Panel: Funders’ Perspectives

Joel Wurl
National Endowment for the Humanities
Washington DC, 20506 USA
jwurl@neh.gov

Joyce Ray
Institute of Museum & Library Services
Washington DC, 20036 USA
jray@imls.gov

Neil Grindley
Joint Information Systems Committee
London, WC2E 7EN UK
callee@email.unc.edu

Kathleen Williams
National Historical Publications & Records Commission
Washington DC, 20408 USA
Kathleen.williams@nara.gov

ABSTRACT
This panel brings together leaders from four funding agencies in the US and the UK -- the National Endowment for the Humanities (NEH) [1]; the Institute of Museum and Library Services (IMLS) [2]; National Historical Publications and Records Commission (NHPRC) [3]; and the Joint Information Systems Committee (JISC) [4]. The panelists will speak on the ways in which their organizations are supporting digital curation initiatives at their respective institutions, and future directions for funding.

Keywords
Digital curation, digital preservation, funding, grant programs.

1. REFERENCES

This work is licensed under the Creative Commons Attribution-Noncommercial-No Derivative Works 3.0 Unported license. You are free to share this work (copy, distribute and transmit) under the following conditions: attribution, non-commercial, and no derivative works. To view a copy of this license, visit http://creativecommons.org/licenses/by-nc-nd/3.0/.

DigCCurr2009, April 1-3, 2009, Chapel Hill, NC, USA
Thinking Like a Digital Curator: Creating Internships in the Cognitive Apprenticeship Model

Elizabeth Yakel
School of Information
University of Michigan
Ann Arbor, MI 48109 USA
yakel@umich.edu

Paul Conway
School of Information
University of Michigan
Ann Arbor, MI 48109 USA
pconway@umich.edu

Magia G. Krause
School of Information
University of Michigan
Ann Arbor, MI 48109 USA
mghetu@umich.edu

ABSTRACT
Effective formal learning about digital curation must take place both in the classroom and in the field. This paper discusses how the cognitive apprenticeship model is being applied in the new Preservation of Information specialization at the University of Michigan School of Information to foster learning inside and outside the classroom. By adopting this approach, our philosophy is that pedagogy is as important as content to achieve the goal of the specialization which is to help students learn to ‘think like a digital curator’ while imparting specific skills. The opportunity to create synergy between courses and internships was made possible by a grant, "Engaging Communities to Foster Internships for Preservation and Digital Curation" from the US Institute for Museum and Library Services.

Keywords
Digital curation, digital preservation, preservation administration, digital curation education, preservation education.

1. INTRODUCTION
Digital curation is a specialization that encompasses both theory and applied knowledge. In formal educational settings, learning to think like a digital curator requires synergy between classroom activities and field experiences. This paper discusses how the cognitive apprenticeship model is being applied in the new Preservation of Information specialization at the University of Michigan School of Information to foster learning both inside and outside the classroom. The overall goal of the specialization is to help students ‘think like a digital curator’ while imparting specific skills and capabilities. The model helps students learn how to learn in the rapidly changing world of digital information. The paper begins with an overview of the Preservation of Information specialization at the University of Michigan School of Information (SI), continues with a discussion of internships and the cognitive apprenticeship model, and then concludes with a discussion of how Michigan is adopting the model to educate a new generation of digital curators.

Although research has been done on the content of digital curation education [11], our model presents a pedagogical framework which we think will complement content. Furthermore, we argue that pedagogy balancing classroom and experiential learning is essential for digital curation education internationally.

2. THE MICHIGAN CONTEXT
2.1 The Preservation of Information Specialization
The SI Preservation of Information (PI) specialization developed in response to an urgent need for expertise in preservation, digital curation, and web archiving in the 21st century. Studies have shown that there is little coursework in these areas and even fewer sustained programs that allow students to focus deeply [7]. Although SI had previously highlighted digital curation issues as minor components of several existing graduate courses (e.g., Electronic Records Management, Preserving Information, Understanding Archives and Records, Archival Appraisal), a digital curation curriculum was not particularly coherent and few internship possibilities existed to reinforce the classroom learning process. Faculty research interests and the increasing expectations of prospective employers opened opportunities for Michigan to address digital preservation and curation issues directly and forcefully. In March 2007, the School of Information and the Rackham School of Graduate Studies jointly approved a proposal to create a new Preservation of Information specialization.

A key assumption of the PI specialization is that preservation administration for analog and digital materials must be approached in a holistic way. Splintering preservation into analog and digital themes is not productive in organizations that need professionals trained and capable of ‘dialog along the analog-digital boundary.’ The generation of preservation leaders currently in place has successfully increased the investment of cultural heritage organizations in preservation. The demand for a new generation of digital curators, however, appears about to explode as organizations with a variety of preservation mandates begin to come to terms with the particular challenge of managing their assets born in or converted to digital form [9].

The PI specialization teaches the principles and knowledge of preservation administration across all media and formats. Areas covered include: collection evaluation and risk assessment, planning, preservation policies, technological infrastructure and standards, facilities and environmental controls, the special
challenges of particular sources, and resources for sustainability and stable funding models.

Students specializing in Preservation of Information must complete 12 credit hours of course work as part of a 48 credit hour Masters’ curriculum. SI 581 – Preserving Information (1.5 credits) is required of all students in the specialization. Students earn the remaining 10.5 credit hours by completing at least seven 1.5 credit (half-semester) courses on specific aspects of preservation (See Table 1). Other courses in the overall SI curriculum augment the specialization. Technologically intense courses, such as “Understanding Networked Computing” and “Database Design,” impart significant knowledge about computing environments. “Collection Development” and the “Digital Libraries and Archives” provide important information about the development of content and researcher services that impact curation. Specialized courses, such as “Understanding Records and Archives,” “Archival Appraisal,” and “Electronic Records Management,” connect preservation thought to the theories of archives and records management. Taken together, the SI Master’s curriculum forms a comprehensive education in the curatorial issues surrounding digital documents.

<table>
<thead>
<tr>
<th>Table 1: Course Overview</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Current Courses</strong></td>
</tr>
<tr>
<td>SI 581: Preserving Information (Required)</td>
</tr>
<tr>
<td>SI 601: Data manipulation</td>
</tr>
<tr>
<td>SI 644: Advanced Preservation Administration</td>
</tr>
<tr>
<td>SI 678: Preserving Sound and Motion</td>
</tr>
<tr>
<td>SI 675: Digitization for Preservation</td>
</tr>
<tr>
<td>SI 639: Web Archiving</td>
</tr>
<tr>
<td><strong>Proposed Courses (Anticipated Start)</strong></td>
</tr>
<tr>
<td>Digital Preservation Management (Winter 2009)</td>
</tr>
<tr>
<td>Physical Treatment Processes for Preservation Administrators (Winter 2009)</td>
</tr>
<tr>
<td>Practical Engagement Workshop in Digital Curation (Winter 2009)</td>
</tr>
<tr>
<td>Science/Social Science Data Curation (2010)</td>
</tr>
<tr>
<td>Economics of Sustainability (2011)</td>
</tr>
</tbody>
</table>

### 2.2 Internships at the School of Information

In designing the specialization, we noted that opportunities for practical engagement with digital curation practices are missing from the overall design, even though the School’s Practical Engagement Program (PEP) is an integral part of the professional Master's program. Thus, full implementation of the preservation specialization was not possible. PEP integrates the application of knowledge and skills to specific problems outside the classroom and requires students to combine what they learn in the classroom with what they observe and experience in the "real world." A structured and required credit-based internship program, PEP provides students with a monitored, mentored, and reflective internship experience that is supported by faculty and the Academic and Career Services Office. PEP is a central activity and a core value of our program. SI requires Master’s students to earn at least six of the 48 credits required for the degree through credit-based PEP internships.

The PI specialization is able to tap a strong support infrastructure for internships, including rules and guidelines to ensure that the internship is an effective learning experience. Mentoring is a critical component of the PEP experience. Students are supervised by an on-site mentor who is far more knowledgeable in the field of work than the student. With the support of SI faculty, mentors and students establish a set of learning objectives prior to the internship and work towards accomplishing those objectives. All mentors meet with their student intern at least once a week to provide guidance and advice and to ensure the student is progressing toward the stated goals and objectives. All internships are approved by the school to ensure that they meet academic criteria:

- Field of work must be something we teach and research at the School of Information
- Graduate-level professional work
- A qualified on-site mentor
- Learning objectives

These guidelines, best practices, and evaluation mechanisms mirror those in the new manual on internships by Bastian and Webber [2] and the evaluation criteria proposed by Bastian. [1].

During the 2007 – 2008 academic year approximately 220 Master’s students participated in PEP-credit internships. In summer 2008, 120 students worked at 140 internship sites across the country and internationally. Students who engaged in these internships contributed to the School’s Online Portfolio System that is monitored by faculty and staff. Students create reports, upload artifacts, blog with other students, and reflect on their own experiences. This system provides particularly essential support to students who are not resident in Ann Arbor during summer internships.

![Figure 1. Screenshot of Online Internship Portfolio System](image)

### 3. COGNITIVE APPRENTICESHIP

In seeking to develop a comprehensive internship component for the PI specialization that could be integrated into the School of Information’s existing practical engagement program, we turned to the cognitive apprenticeship model [3], [4] for guidance and inspiration. The primary motivation of the cognitive apprenticeship model is the desire to translate the reasoning and strategies that experts employ to solve complex problems into a student learning environment. The learning model focuses on intellectual activities, as opposed to physical tasks, and is therefore the most appropriate approach when knowledge...
acquisition is based on observable skills. The field of digital curation fits well into the cognitive apprenticeship model because successful curation is as much about thinking through and analyzing preservation problems as it is about actually accomplishing a physical task. For example, in web archiving, establishing the parameters of a crawl and making selection decisions has as greater impact than setting up the technology to actually accomplish the crawl.

Collins, Brown, and Holm [5] propose that transparency of work process is a vital aspect of apprenticeship. To make work visible, they propose a model of cognitive apprenticeship based on four dimensions: content, method, sequencing, and sociology.

3.1 Content
Content combines all the types of knowledge that comprise expertise in a given subject area. This includes declarative knowledge consisting of the definitions, concepts, facts, and techniques experts use to accomplish tasks, “metacognitive” or procedural knowledge used to analyze and manage tasks, and the strategies that experts use to learn new ways to accomplish tasks.

3.2 Method
Method refers to the pedagogies and instructional designs that best allow learners to gain the knowledge and skills to develop and extend expertise in a subject. Three methods (modeling, coaching, and scaffolding) focus on the role of the teacher or mentor. Another three methods (articulation, reflection, exploration) focus on the role of the learner.

- In modeling, an instructor or mentor verbalizes tacit knowledge by working through the process of accomplishing a task in such a way as to make her thinking clear to students.
- Coaching encompasses mentoring learners through tasks by offering guidance and providing feedback.
- According to Sawyer [13: 11] scaffolding is “the help given to a learner that is tailored to that learner’s needs in achieving his or her goals of the moment.” Scaffolding can be verbal or physical (such as demonstrating some technique) and is usually accompanied by a gradual reduction of support so that a student begins to accomplish tasks on his or her own.
- For a student, articulation is the verbalization of his or her thinking and rationale for actions.
- Reflection is the process by which a student compares work with someone else’s, another student or an expert.
- Exploration encourages a learner to solve problems alone and independently.

3.3 Sequencing
Sequencing assists students to create mental maps based on a framework for knowledge acquisition. According to How People Learn [8], there are three elements in sequencing:

- Increasing complexity builds students’ expertise by presenting them with more and more difficult tasks or problems to solve.
- Increasing diversity encourages students to utilize a wider variety of tactics in their work.

Global to local sequencing helps students gain a big picture of the task at the onset to better enable them to create a localized conceptual understanding of a specific process. When this is successful, students can transfer this conceptual model to small and / or related tasks in the future.

3.4 Sociology
Sociology: the fourth dimension of the cognitive apprenticeship framework, emphasizes the sociocultural influences on learning [14]. Learning is situated in social, cultural, and technological contexts, what Collins [4: 52] calls the “sociology of learning.” In situated learning [3], [10] students are encouraged to engage in learning that has real-world applications.

3.5 Applying Cognitive Apprenticeship
With the pending implementation of a broad practical engagement program, the cognitive apprenticeship model is now a critical pedagogical driver for the entire Preservation of Information curriculum at SI. Classroom assignments prepare students to work in the field for a term or a summer. With the support of the U.S. Institute for Museum and Library Services, SI is building an educational environment that explicitly creates synergy between formal coursework and practical engagement experiences. Funding for the project, "Engaging Communities to Foster Internships for Preservation and Digital Curation," supports two types of internships: 1) ten 360 hour summer internships at organizations around the country that possess digital curation expertise; and 2) a practical engagement workshop course that incorporates opportunities for semester-long internships with local organizations where faculty will co-mentor students working in organizations with less expertise in digital curation.

3.5.1 Course Assignments
In terms of the actual courses, we create synergy between the classroom and real world sites. This involves placing students in situ and working on assignments that come up naturally in the workplace, as well as introducing activities more abstractly in class that are set in realistic contexts that make sense to students. This tight coupling between course work and ‘real life’ work is one of the core values in the cognitive apprenticeship model. Table 2 lists selected PI courses and major assignments that require students to perform some task in the field.

<table>
<thead>
<tr>
<th>Course Assignment Example</th>
<th>SI 581 Preserving Information</th>
<th>Preservation Assessment</th>
</tr>
</thead>
<tbody>
<tr>
<td>SI 644 Advanced Preservation</td>
<td>Preservation Plan</td>
<td></td>
</tr>
<tr>
<td>SI 675 Digitization for Preservation</td>
<td>Grant Proposal for Digitization</td>
<td></td>
</tr>
<tr>
<td>SI 639 Web Archiving</td>
<td>Design and Execute a Crawl</td>
<td></td>
</tr>
<tr>
<td>Digital Preservation Management</td>
<td>TRAC or DRAMBORA Assessment</td>
<td></td>
</tr>
</tbody>
</table>

Assignments are sequenced. Students must complete SI 581 Preserving Information first; the other classes build on this course’s foundation. Sequenced learning also helps students begin to develop a framework for problem solving through the mastery of tangible and realistic workplace skills. This enables later transfer or application of that knowledge in other, perhaps
students who are also doing internships. Students contribute to a shared discussion forum with other students in internships, helping to bridge what is often significant physical distance and a reduced opportunity for reflection. Summer interns, along with all SI students who participate in a summer PEP internship, are required to utilize the Online Portfolio System. This is a reflective learning tool. Through the online portfolios, students reflect in both the short and long-term on their internship experience by submitting bi-weekly reports, artifacts, and a summative assessment of the learning goal and objectives. The online portfolio is also a way for SI to stay connected with the students and their mentors as the internships progress. Faculty and SI Career Services personnel track work progress, hours worked, and stay on top of any issues or problems the student may encounter. Students in internships also participate in discussions or blogs with their peers using the system which enhances learning through reflection. In the portfolio system, students contribute to a shared discussion forum with other students who are also doing internships.

3.5.3 Summer Internships in Digital Curation
SI has partnered with ten organizations to offer summer digital curation internships for three years, beginning in the summer of 2009. The ten sites for the summer internships are: the Center for Research Libraries (CRL), the Florida Center for Library Automation (FCLA), the Internet Archive, and the LOCKSS program at Stanford University, the Northeast Document Conservation Center (NEDCC), OCLC, the Smithsonian Institution Archives, Safe Sound Archive, the Inter-university Consortium for Social and Political Research (ICPSR), and the University of Michigan Libraries. Partnering sites were chosen for three reasons: their track record for success in some aspect of digital curation, they have staff capable of and willing to provide formal mentoring of interns, and cumulatively they represent a diversity of types of organizations and expertise. Each of the internship sites has developed a position description for one or more specific projects. Student, mentor, and faculty co-develop learning objectives that benefit the student and contribute to the work of the host institution. The summer internships feature students doing a variety of digital curation activities from the very technical application of specific software applications, to policy development and research, to field work and client service. Each of the internships is paid and requires a minimum of 360 hours of work; students receive 6 hours of academic credit for successful completion of the internship.

In designing the summer digital curation internships, we carefully selected mentors at each site who can articulate tacit knowledge and rational for actions. Making work visible, an important aspect of cognitive apprenticeship is an element in all the internships, whether it is learning the process of carrying out observable tasks or situating students in internships with mentors who deliberately bring thinking to the surface.

3.5.4 Term-Based Practical Engagement

In conjunction with the practical engagement training course, we have recruited local organizations for digital curation internships. This component is designed to extend opportunities for students to participate in digital curation / preservation administration internships who are not willing or able to travel for a summer internship. Students participating in local internships will earn 1.5 credit hours and be expected to contribute 90 hours of volunteer labor to a single organization. We anticipate providing at least 15 local internships each year, half of them in departments or units of the University of Michigan and half in nearby cultural organizations.

3.1 We anticipate that some of the mentors at local internship sites will need assistance with the digital curation assignments or and even mentoring services to design projects, create appropriate learning objectives, and supervise the interns. The practical engagement training course is designed so that the instructor can act as a co-mentor for local projects, if needed, to ensure that both the student and the host institution benefit. This type of arrangement greatly expands the local sites available as well as provides unique benefits for sites with limited digital curation expertise. For example, this type of arrangement would allow students to complete digital readiness surveys or potentially a Trustworthy Repositories Audit & Certification (TRAC) TRAC [12] or DRAMBORA [6] audit in a wide variety of institutions or enable web archiving projects in diverse institutions.

4. CONCLUSIONS

Getting students to ‘think like a digital curator’ in a professional community where few individuals grasp these concepts is very challenging. The School of Information has designed a preservation/digital curation program to leverage existing expertise, adopt a cognitive apprenticeship model in an integrated classroom-fieldwork environment, and support students through both formal course work and faculty mentoring. This approach considers pedagogy as well as content as essential components in of the curriculum. We also think that this approach is portable and would work in a variety of education programs internationally. Digital curation is at a transition point. At the moment when more and more individuals and organizations are acknowledging its importance, there are relatively few individuals employable at the entry professional level and capable of addressing the technical and administrative challenges of digital information. Educating a new, dare we say first, generation of digital curation specialists is essential, but education must be done in a way that creates a
group of learners who are prepared to grow and learn in a rapidly evolving digital universe.

5. ACKNOWLEDGMENTS
The authors wish to thank the Institute for Museum and Library Services for funding this project (RE-01-08-031) that incorporates the development of a practical engagement program.

6. REFERENCES


Qualification & Education in Digital Curation: the nestor Experience in Germany

Heike Neuroth
Göttingen State and University Library
37073 Göttingen, Germany
neuroth@sub.uni-goettingen.de

Achim Osswald
Cologne University of Applied Sciences
Institute of Information Science
50678 Cologne, Germany
Achim.Osswald@fh-koeln.de

Stefan Strathmann
Göttingen State and University Library
37073 Göttingen, Germany
strathmann@sub.uni-goettingen.de

ABSTRACT
Being a relatively new topic in research and education, digital curation is, for a number of reasons, currently not very well covered by university curricula. Within the project "nestor", a transnational partnership of academic institutions in Germany, Switzerland, and Austria, has established a comprehensive qualification program not only based on e-learning tutorials but also on schools, seminars and an (open access) encyclopedia in digital curation.

Categories and Subject Descriptors

Keywords
Digital long term preservation, digital curation, education, further education, nestor, qualification.

1. DIGITAL CURATION: A CHALLENGE FOR EVERY MODERN SOCIETY
In 2006, German experts, institutions, and societies agreed the principles in “Memorandum on the long-term accessibility of digital information in Germany”; abbreviated as "nestor – memorandum”1 which provides a political state of the art overview and offers perspectives for further activities on digital curation in Germany:

“Digital information has become an integral part of our cultural and scientific heritage. We are increasingly confronted with scientific findings, historical events and cultural achievements presented in electronic form. (…)

Preservation of the digital heritage requires additional and sustained effort on the part of the policy makers, authors, publishers, hard and software manufacturers, and the cultural and scientific memory organisations.

The necessary framework conditions for this should be anchored in the national long-term preservation policy. This should contain intelligent strategies for preserving the knowledge manifested in digital objects, and recommend measures for their long-term preservation and accessibility in the form of a sustainable science and research policy.”2

Germany - like other countries in the EU facing this challenge - has created an umbrella of projects and activities on long-term preservation. Qualification issues are becoming an integral part of this umbrella.

2. EDUCATION AND QUALIFICATION CHALLENGES IN GERMANY
Born digital objects as well as those that are retrodigitized are a relevant part of collections in institutions such as libraries, archives and museums. These cultural heritage institutions have come to the attention of the EU, the German Federal Ministry of Research and Science, the German Research Foundation and other organizations in Germany.

The virtual headquarters of German activities is the website of the well-known German project nestor (nestor - Network of Expertise in Long-Term Storage of Digital Resources)3.

The nestor project is two-phased. Its first phase (2003-2006) resulted in several basic documents such as the Memorandum cited above. Its second phase (2006-2009) is based on several work packages realized by working groups (WG).

These are

1See the text at

2 ibid.

3 See: http://www.digitalpreservation.de
The nestor-memorandum is a future oriented and promising framework for the preservation of the digital cultural heritage in Germany. In paragraph 18, the nestor-memorandum focuses on qualifications.

18. Long-term digital preservation generates new tasks for the preserving institutions. Professionally trained staff must be used. The requirements and tasks of long-term digital preservation need to be given appropriate attention in the initial and further training programmes. Specific training courses need to be set up which sensitize their participants to specific topic areas and which also qualify them to perform the tasks at hand.

Göttingen State and University Library (SUB) initiated a working group “Cooperation with Universities in qualification” when the second project period of nestor was begun. The WG brings together participants from several cooperating universities as well as other institutions involved in the qualification of information professionals. Meanwhile, the cooperation has been expanded to institutions in Switzerland and Austria, adding further experience and insights from those German speaking countries. The partners in this cooperation are

- Göttingen State and University Library
- Cologne University of Applied Sciences, Faculty of Information and Communication Sciences
- Humboldt-Universität zu Berlin - Faculty of Arts - Institute of Library and Information Science
- Leipzig University of Applied Sciences, Department of Media and Communication
- University of Applied Sciences Marburg (Archivschule Marburg)
- University of Applied Sciences Potsdam, Department of Information Science
- University of Applied Sciences Chur, Switzerland, Information Science

Although they have different foci, the partners have in common their willingness and commitment to improving the qualification of people involved in digital curation.

3. UNIVERSITY BASED COURSES - A PATCHWORK

3.1 State of the art

A survey on qualification in digital curation at universities must start at the Library and Information Science (LIS) related departments / institutes of universities.

In 2006, Osswald and Scheffel [Osswald/Scheffel 2007] surveyed Bachelor- and Master's programs offered or planned by LIS departments / institutes in the German speaking countries Germany, Austria and Switzerland. In total 16 universities in Berlin, Darmstadt, Düsseldorf, Hamburg, Hannover, Hildesheim, Cologne (2x), Constance, Leipzig, Potsdam, Regensburg, Stuttgart (2x), Chur (CH), Eisenstadt (AU) and Krems (AU) were surveyed.

Teaching staff from each institution complained that none was able to cover the topic completely or to focus on current developments. Specially appointed professors for the field haven’t been available and there was (and still is) little likelihood that the entire topic will be assigned to one professorship.

As a consequence, subject-related specialization of those academics involved results in a focus on special topics within the field of digital curation. This provides a good starting point for cooperation among different universities.

3.2 Education & Qualification Programs

3.2.1 Bachelor-programs

Several of the LIS-related BA programs offer various courses related to digital curation. Most of them concentrate on a basic introduction to this field. Some of them take the opportunity to go into more detail in more specialized courses related to preservation or information technology.

Those nine universities that indicated that they deal with the topic focus on it in between 2 and 10 class sessions (each 90 minutes) or devote a whole course to it – many in connection with topics like information management, records management, digital/electronic publishing, archival science or documentation activities in museums.

Topics mentioned in these class sessions / courses:

- Long term preservation of digital data
- Metadata
- Archival repositories / Open Archival Information System

10 See: http://www.fh-htwchur.ch/Information-Science.17.0.html?&L=1

11 See: http://www ifs.tuwien.ac.at/dp/
• Persistent identifiers
• Formats
• Open Access
• Legal issues
• Data protection and archiving.

Within several of the courses, parts of the IMARK (Information Management Resource Kit) modules of UNESCO and/or the Tutorial on Digital Preservation Management of Cornell University have been used – besides other materials and exercises developed by the lecturers themselves.

3.2.2 Master-programs

For obvious reasons, there is a trend to realize an in-depth qualification in digital curation within a MA program. This offers the opportunity to combine existing knowledge in a specific field (subject-related bachelor's degree or a LIS-, archival- or museum-related bachelor's degree) with preservation know-how. Such an additional job-related qualification might someday be an obligatory component of management positions.

It is also very likely that such a qualification will be acquired within continuing professional education on a part-time basis. Future developments will show which qualification track will be most suitable to the needs of practitioners.12

A program which touches the topic but does not focus only on digital preservation is the four semester MA-program “Conservation of New Media and Digital Information” of the State Academy of Art and Design Stuttgart (Staatliche Akademie der Bildenden Künste Stuttgart)13 which is designed for members of institutions dealing with modern art. Its objective is:

Imparting knowledge and capabilities about long-term preservation of fine art, cultural, archival and library heritage in the fields of photography, video und digital information.14

4. QUALIFICATION NEEDS

Long term preservation of printed and digital objects has been a topic in educational activities at universities for a number of years. Starting with courses in further education related to durability of digital media like floppy disks, CD-ROMs and now DVDs, there is a growing awareness that there are other critical areas as well: availability of application software, operating systems and know-how as well as the persistence of formats, etc. These topics have been picked up within courses at university programs, especially by those dealing with technical aspects of applications in the field of library science, archival science and museology. However, aside from initiatives by some information science departments, no systematic qualification focusing on the topic has taken place up to now.

Therefore, the average librarian or member of an archive has a lack of practical and conceptual knowledge about the topic. Colleagues involved in the nestor project have tried to fill this gap by offering one-day workshops or seminars giving a comprehensive view on the challenges of digital curation. They recorded their presentations and distributed them free of charge on DVD. Nevertheless, the number of practitioners reached was not satisfactory. It was felt that the experts on specific fields could not be hired again and again for events such as this.

When the first period of the nestor project was completed, a workshop was organized by the German National Library in June 2006 titled “Den Fortschritt bewahren” (Keeping progress).15 One of the presentations focused on the topic “Learning and transfer: Education and further education in the field of long term preservation”16. It emphasized the need for a systematic approach to the qualification issue. The authors suggested that courses in the field of digital curation should be differentiated in regard to the target groups as well as to the usage of media and the didactic concept related to it.

After analyzing the situation, it was suggested that at least five target groups should be taken into focus:

• the management of institutions to which curation is or will be a relevant topic by spreading awareness of the challenge to be faced
• middle-management in cultural heritage institutions by providing them state-of-the-art information on concepts and methods of digital curation
• staff members of cultural heritage institutions who are involved in or who are organizing curation activities on an operational level
• BA-students in programs qualifying for jobs in cultural heritage institutions
• MA-students in programs qualifying for projects and management jobs in cultural heritage institutions.

12 E.g. in the MA program "European Multimedia, Arts and Cultural Heritage Studies" provided cooperatively by the Universities Cologne (Germany; Prof. Thaller), Coimbra (Portugal), Lecce (Italy) and Turku (Finland) digital curation is supposed to be a topic on which students can stress on. See more at http://www.euromachs.net/.

13 See for details http://www.mediaconservation.abk-stuttgart.de/home_e.htm

14 http://www.mediaconservation.abk-stuttgart.de/home_e-Dateien/E_allg.htm. This highly subsidized application oriented program asks for a tuition fee of about 1500 € per term and is very exclusive regarding the number of participants: 12 per year! Experts are hired for presentations and courses (or parts of them) because the broad spectrum of specific know-how is not available by the regular teaching staff. The programme will be evaluated in 2009 and it’s continuation is not guaranteed.

15 See the presentations at http://www.langzeitarchivierung.de/modules.php?op=modload &name=Downloads&file=index&req=viewsdownload&sid=20

5. PROJECT BASED ACTIVITIES

At present there are at least five lines of activities in digital curation-related qualification of the nestor WG “Cooperation with Universities in qualification”:

- nestor seminars
- nestor Schools
- the nestor Handbook
- potential development of a cooperative curriculum
- development of e-learning tutorials

5.1 nestor-seminars

The first project period of nestor had no special focus on qualification issues. Nevertheless, those active in nestor realized that there was a need to spread the research results to a community of people interested or involved in the topic. Consequently the results, reports and presentations of the nestor project have been made public. But it seems that only the free provision of information is not the solution of the problem.

Additionally nestor-related researchers set up an activity line called “nestor seminars” (nestor Seminare). Two of them have been recorded on video and distributed on DVD free of charge [SUB / Neuroth 2006]. The idea was to multiply knowledge via different means of distribution – including visualization and some personalization (which is provided by seeing the people presenting their slides and speeches on video).

Beside these seminars, the experts gave presentations at conferences and also within university programs – but this could not be done on a systematic basis.

Nevertheless, occasional nestor-related seminars are taking place. They focus on special topics and audiences. For instance, in May 2007, the German National Archive offered a one-day-seminar on the topic of “The role of Open Archival Information System (OAIS) in setting up a digital archive: A helpful guideline for German archives or pure theory?”

As mentioned earlier, this kind of distribution of knowledge is limited especially as to the time that the experts have available. The planning of nestor phase 2 offered the chance to put more concentrated emphasis on the qualification issue.

As a consequence, the nestor-School concept has been further developed. It is influenced by good experiences in DPE and DELOS summer schools and the practical training sessions there. DPE and DELOS are both European projects with a strong focus of qualification and education in digital curation.

Complementary for all those not being able to take part in the School events a Handbook-concept was implemented.

5.2 The nestor Schools

On its qualification website nestor reports:

“There will be different events to support the training and education needs. These training events have started with the nestor Spring School 2007 and are continued with workshops and seminars. These ongoing series of events is intended to serve the needs of different communities.”

The nestor Spring School took place in March 2007 in a small and remote village in the eastern part of Germany. The stimulating atmosphere of a former monastery offered an exclusive chance to the about 40 participants to learn more about preservation of digital objects.

nestor Spring School 2007 had the topic “Introduction to digital long term preservation” and lasted five days. Its objective was a basic introduction to the topic, discussion and presentation of solutions and practical training lessons in small groups. Additionally the purpose of the School was the exchange of experiences and networking between different application areas.

Slides, protocols and results of the School-event are made available to the participants on a website.

The event was inspired and supported financially by Digital Preservation Europe (DPE) and DELOS, the “Network of Excellence on Digital Libraries” which is funded by the European Commission in the frame of the Information Society Technologies Programme (IST).

Participants received a certificate signed by the organizers and lectures of the School-event mentioning the credits assigned to it. These credits – based on the European Credit Transfer System (ECTS) have been accepted for regular curricula (BA and MA) e.g. of University of Applied Sciences Potsdam or Cologne University of Applied Sciences.

15

21 See e.g. http://www.dpc.delos.info/ss07/
22 http://nestor.sub.uni-goettingen.de/education/index.php?lang=en
23 See for details http://www.digitalpreservationeurope.eu/
Here are some selected details of the topics of presentations and exercises during the Spring School 2007:

- Introduction to digital long-term preservation
- Metadata: What’s it? Why needed? How to be applied?
- Format: An important metadata element. (Significant Properties, File Format Registries, Tools …)
- OAIS as a model for digital curation
- Criteria for the selection of digital objects to be preserved
- Basic parameters for Preservation Policy
- Trusted repositories

The success and stimulating responses of the participants have motivated the organizers to continue with their efforts. The nestor Spring School 2007 was followed by two similar designed seminars: “The nestor Winter School” 2007 at the same location which focused on practical application aspect of digital curation and – in 2008 – “The nestor Summer School” which took place in Staufen/Breisgau. The next nestor School-event is scheduled for March 2009. All this school events take place in cooperation with and supported financially by DPE. On the other hand nestor cooperates with the schools organized by DPE (2007 Lithuania, 2008 Czech Republic, 2009 Spain).

5.3 The nestor Handbook

An ongoing nestor project activity is the publication of “The nestor Handbook” which is based on contributions of experts in the field of digital curation and was published as version 0.1 in March 2007. It is revised regularly (recent version 1.5). An additional printed version 2.0 is planned for summer 2009. The Handbook called “nestor Manual - A compact encyclopedia of long-term digital preservation (Eine kleine Enzyklopädie der digitalen Langzeitarchivierung)” [Neuroth et. al 2007f.] is characterized on the nestor-website like this:

“The nestor-manual <=Handbook; Ed. > tries to bundle the recent state of knowledge on digital long-term preservation and its various components. Collected and compiled in this "small encyclopedia" the authors give the German speaking community access to the subject in a structured manual.”

Some of the chapter titles are:
- State-of-the art of legal aspects
- National Preservation Policy
- Institutional Preservation Policy
- The OAIS (Open Archival Information System) reference model
- Trusted repositories
- Formats
- Relevant standards important in digital curation (e.g. PREMIS)
- Strategies of digital preservation

The digital version is provided free of charge under a Creative Commons License and available for download on a chapter basis or as a complete document (up to now more than 350 pages).

5.4 Development of a cooperative curriculum

Regarding the objectives of the nestor working group “Cooperation with Universities in qualification” it has never been the intention to set up an own digital curation curriculum. This would have been in contradiction to the funding Ministry which is a federal body. But nestor has the objective to stimulate the development and realization of such a program – whether it will be realized as further education or continuing professional education within a MA program.

Based on the understanding that none of the partner’s active in the nestor working group will be able to set up a preservation-related curriculum by its own, it was decided to do this on a cooperative basis. This received support by experts and the administration of several universities in Germany and in other German speaking countries. Meanwhile a “Memorandum of understanding” has been signed by ten partner institutions to cooperate in developing and adjusting modules focused on selected topics of preservation. They will be provided by the cooperating parties. Courses of other universities will be accepted in a local curriculum. The workload of each module will be calculated in credits in regard to the ECTS-system (European Credit Transfer System).

5.5 Development of e-learning modules and other plans within the nestor project

Meanwhile a main part of the partner’s activities is the development of e-learning tutorials on digital preservation.

This is an opportunity to cooperate with other initiatives in European and other countries. And it is a chance to provide the content of the handbook and the presentations in English and connect them with didactically empowered content in e-learning modules.

During 2007/2008 e-learning modules are developed within student-based projects at the Universities of Applied Sciences at Chur (Switzerland), Cologne, Leipzig and Potsdam. Based on the e-learning platform moodle which is hosted by the Computer and Media Service (CMS) of Humboldt-Universität zu Berlin several courses in curation-related topics have been developed:

- The OAIS model
- Formats
- Compilation of a Submission Information Package (SIP)
- Significant Properties
- Digital Preservation of CAD-data
During winter term 2008 several additional courses will be
developed and the existing ones will be enhanced. Quality
management of the courses available and getting developed offers
new chances to the Universities involved to intensify their
cooperation and form strategic cooperation.

A survey regarding quality criteria of e-learning material showed
clearly that there are no agreed standardized criteria to measure
the quality of e-learning tutorials [Osswald/Otto/Stettler 2009]. The
quality criteria chosen by nestor to be in focus are

- **technical** like stability, ability to run on different
  platforms, performance, and relation to standards
- **didactical**: stimulating self-paced and self-controlled
  learning and controlling the level of knowledge reached
- **usability** related (including understandability)
- **content** related regarding validity, integrity.

Based on a cooperative activity there will be several steps taken
to improve these aspects of quality of the courses

- **before** the production process by enhancing the
  awareness of the didactical and structural concept of the
  courses and the tools and steps relevant to apply them;
- **during** the production process by applying collaborative
  assessment on the results available by checking them
  reciprocal in relation to the overall goals given;
- **after** the production process by evaluating the tutorials
  by different student groups of the Universities being
  active in the cooperation.

6. **SYNERGY BY INTERACTION BETWEEN UNIVERSITY AND PROJECT
BASED QUALIFICATION**

There is some evidence that the activities of the nestor project and
mainly the working group “Cooperation with universities in
qualification” has stimulated reflections and awareness of digital
curation even in those universities where the topic has been part
of courses or programs since years. The results of nestor have
influenced the content of lessons and courses provided. But nestor
has also initiated cooperation where competition seemed to
dominate: Instead of trying to hire a person out of the expert pool
of nestor to offer specialized courses at single institution
cooperation is planned.

The intra- and cross-sectoral cooperation offers new opportunities
for participation in the results each party has available to all those
who are willing to cooperate. nestor as an umbrella additionally
offers the option to receive support by further national and
international partners e.g. like DPE. nestor finally is a guarantee
for a qualitative level of content related to those modules which
will be certified by nestor – both in regard to research findings of
the project in its different stages and didactically. nestor brings in
the authority of a project in which leading German cultural
heritage institutions are playing the major role. This is the
network of competence and knowledge to which all partners are
linked.

On the other side nestor as well receives insights and links to
universities and qualification bodies which have never been
reached in such a way before.

7. **OPEN ISSUES**

Nevertheless the first steps are undertaken and initiated by the
nestor working group some other issues came up. The concept
“developing e-learning tutorials by students for students” means
that only in combination with an university lessons, modules
could be developed. The consequences are that the development
of e-learning tutorials is very slowly, because they must be
integrated in the regular curriculum and there must be an interest
from student side in developing such courses. On the other side
the initial input given by the teachers by starting these courses
every winter or summer time is high, because every new student
group has to be introduced in the policy of developing such
tutorials and into the guidelines.

The advantage is that the acceptance of using these tutorials is
high. The students are not only getting skills in the field of digital
curation but also in developing e-learning materials and learning
in small project groups (soft skills).

Another open issue is related to quality management of these e-
learning tutorials. Are the measures taken to ensure the quality
of the tutorials is effective? Will it be helpful to establish an external
“evaluation board” and what should be evaluated them?

All in all, the described activities are a very promising way to put
the attention on the need of having qualification and education
methods in place. Without putting enough funding money in this
field we will not be able to preserve our cultural heritage not to
speak from digital curation of research data, which seems to
become more and more important (see e.g. ongoing international
discussions on “open data”).

It is not only a big challenge but also a big chance to work
together on a university level, this is in the described form unique
in Germany and German-speaking countries. It would be great if
politicians and funding agencies could support this strong
cooperation regardless of local or national structures.

8. **REFERENCES**

[1] Borghoff, Uwe M., Peter Rödig, Jan Scheffczyk: Long-Term
Preservation of Digital Documents. Principles and Practices,
Berlin 2006

information in Germany, Frankfurt 2006
http://www.langzeitarchivierung.de/downloads/memo2006-
e.pdf

curricularer Module zur digitalen Langzeitarchivierung im
Rahmen des nestor II Arbeitspaketes 5. Göttingen 2007
http://nestor.sub.uni-goettingen.de/education/mou.pdf

[4] Neuroth, Heike; Liegmann, Hans; Oßwald, Achim; Scheffel,
Regine; Jehn, Mathias, Strathmann, Stefan (Ed.): nestor
Handbuch: Eine kleine Enzyklopädie der digitalen
Langzeitarchivierung, Göttingen 2007 et seq.
http://nestor.sub.uni-goettingen.de/handbuch/index.php


Educating Archivists about Copyright: How Can We Do It Better?

Jean Dryden
College of Information Studies
University of Maryland
College Park, MD 20742 USA
Jdryden@umd.edu

ABSTRACT
A recent study that looked at the sources and quality of Canadian archivists’ copyright knowledge in relation to repository practices in digitizing archival holdings for Internet access found that Canadian archivists get their copyright knowledge from a range of sources, not all of which are authoritative or up-to-date. Consequently, copyright issues are not always addressed correctly, with adverse effects on access to, and use of, digital resources. If a good knowledge of copyright is an essential component of digital curation, this study suggests that there are some problems in copyright education, and proposes possible ways of addressing them.

Keywords
Copyright, copyright education, digitization--copyright aspects, digital curation--copyright issues.

1. INTRODUCTION
Digital and communication technologies provide exciting new opportunities to increase access to cultural heritage materials that traditionally have been available only in the library, archives or museum that preserves them. Attracted by the possibility of world-wide access to such resources, cultural heritage institutions have enthusiastically begun to digitize their holdings in order to make them available via the Internet. However, such institutions face a host of new challenges as they enter the digital domain. Among these is copyright. Copyright law is a policy mechanism that provides limited monopoly rights to creators as an incentive to disclose and publish works in order to “[serve] the interests of further creation and the growth of knowledge and culture” [1]. In doing so, the law attempts to balance a complex array of competing private and societal interests, including those of creators, copyright owners, a wide range of users, and institutions that preserve protected material in order to make it available for use.

Copyright law and cultural heritage institutions have a goal in common: to make cultural heritage material available in the long-term interest of society. However, Zorich’s survey of digital cultural heritage initiatives refers to “the online intellectual property miasma” [2], and manuals on the successful conduct of digitization projects suggest that copyright is widely perceived to be a barrier to making the holdings of cultural heritage institutions available on the Internet due to difficulties in obtaining permissions, and uncertainty about the application of copyright in the digital environment [3]. It is not surprising that practitioners approach copyright issues cautiously, basing their decisions on risk assessment [4]. Additionally, the need to manage and preserve digital objects over time is an area of pressing concern. Although no clear best practices have yet emerged, the digital preservation methods that are currently available involve a number of copyright issues [5]. All this suggests that an understanding of copyright law is an essential component in preparing students for successful careers in digital curation.

2. THE STUDY
If knowledge of copyright is an important part of digital curation practice, a useful starting point for curriculum development is to see what those involved in digitization projects already know about copyright, and where they get their copyright knowledge. This paper presents the findings of this aspect of an exploratory study of the copyright practices of Canadian archival repositories in digitizing their holdings for Internet access [6]. Although the study looked primarily at repository practices, it is staff members who learn about copyright, communicate that knowledge to their co-workers, and transform it into institutional policies and procedures. Thus, the study assumed that institutional practice is influenced by the sources of archivists’ copyright knowledge and the quality of that knowledge. In order to explore those aspects, the study asked the following research questions:

- Where do Canadian archivists get their knowledge of copyright?
- What do Canadian archivists know about copyright?
- How do Canadian archivists perceive copyright?

The study employed the following sources of evidence: the website content of 154 Canadian repositories whose websites feature archival material from the repository’s holdings; copyright policy and procedure documents of those repositories; 106 responses to a questionnaire sent to the staff of those repositories; and 22 interviews with repository staff members.

3. SOURCES OF ARCHIVISTS’ COPYRIGHT KNOWLEDGE
How archivists learn about copyright, and the sources of their knowledge influence repository practices. The study looked at different aspects of archivists’ sources of copyright knowledge, including how archivists learn about copyright initially, how they keep up-to-date, where they seek answers to specific questions,
and how they learn about repository copyright practices. In order to explore how archivists learn about copyright initially, questionnaire respondents were asked, “Where do you get your knowledge of copyright?” and were given a list of twelve ways that archivists could learn about copyright (as well as space to add others) and asked to mark all that applied. Four sources (workshops sponsored by professional associations, colleagues in the profession, books and newsletters, and the statute and regulations) were marked by at least half of the respondents. The top two were professional association workshops and colleagues in the profession, mentioned by 88 (83%) and 85 (80%) of the 106 respondents respectively. Books and newsletters followed closely, marked by 76 respondents (72%). Sixty-five respondents (61%) marked statute and regulations. The remaining sources were marked by fewer than half of the respondents.

Respondents were then asked to rank the top three sources of their copyright knowledge in order of importance. Responses were coded so that a ranking of 1 (most important) was given a value of 3, a 2nd place ranking was given a value of 2, and a 3rd place ranking a value of 1. Workshops sponsored by professional associations (which received a score of 155) were considered by far the most important source of copyright information, followed by books and newsletters tied with the statute and regulations (score 79), and colleagues in the profession (score 72).

None of the questionnaire respondents reported learning about copyright in the course of their graduate education, although one interviewee mentioned that she had been exposed to copyright as part of her graduate archival studies program. This may not be surprising, given that 73% of the 106 questionnaire respondents have been employed in archives or related institutions for more than a decade, and presumably have been out of school for some time. Furthermore, the extent to which copyright has been addressed in graduate programs is not known, although a preliminary examination of the courses and syllabi currently available on the websites of Canadian LIS programs reveals little explicit coverage of copyright.

The questionnaire responses provided a direct answer to the question, “where do you get your knowledge of copyright?” However, policy documents available on repository websites or submitted with questionnaire responses provide indirect evidence of the print and online sources of Canadian archivists’ copyright knowledge. Five repository policy documents, intended primarily for internal use, contain footnotes or bibliographies that document the secondary sources on which the policy was based. A number of these sources are out of date in that they do not reflect the most recent (2004) amendments to the statute. Twelve repositories, in online documents primarily intended for users, direct their researchers to sources of further information about copyright. Such sources are usually links to the statute or to websites of federal government departments or agencies with a mandate for copyright matters.

Learning about copyright once is not sufficient; it is also important to keep up-to-date. Copyright is an evolving area of law, particularly in Canada, where the federal government launched an ongoing copyright reform initiative in 2001, and where the Supreme Court has, over the past five years, issued several landmark decisions relating to the application of copyright in both the analog and digital environments. Questionnaire respondents were asked, in an open-ended question, how they keep up-to-date with changes in the law.

Three noted that it was not easy to keep their knowledge of copyright current. Just over a third of respondents (35%) mentioned only one means of keeping up-to-date; 65% of respondents mentioned two or more. Listservs, mentioned by 42 respondents, are by far the most common means of keeping current. The second most common means of keeping up-to-date (mentioned by 34 respondents) is publications, a category that included books, newsletters, articles, and printed material. Workshops and training activities were mentioned as often as colleagues in the profession (each being mentioned by 20 respondents) as a means of keeping up-to-date. Keeping current through colleagues included emails from, or discussions with, colleagues both in the same repository and in the wider profession. Following closely (18 respondents) was information or notices from professional associations. Other means included Internet sources (8), legislation (7), professional meetings (5), legal counsel (4), and news media reports (3).

How archivists keep up-to-date with changes to copyright law was also explored in the interviews. Interviewees mentioned two particular ways of keeping current about copyright: the national listserv and workshops. Seven mentioned the messages posted on the national listserv by the Copyright Committee of the national professional association as the prime way in which they become aware of changes to copyright law. Two more stressed the importance of workshops as a means of learning, not just about copyright generally, but also about amendments to the law. Three mentioned the role of the Copyright Committee as the body responsible for informing the archival community of changes to the law.

Archivists’ awareness of recent (2004) changes of particular relevance to archivists (statutory amendments and a Supreme Court decision that clarified fair dealing) was explored in the interviews. Five interviewees noted that they keep a reference file on copyright to which they would add any information disseminated on the listserv (or from other sources) regarding recent developments. While few could accurately recall the details about the specific events, they noted that, had these questions come up in the course of the workday, they would have gone to their reference file. Two others remarked that they would ask a particular colleague that they rely upon for copyright information. They appear to rely on others (particularly the national professional association Copyright Committee) to inform the community of changes once they actually occur.

1 Following the 1997 amendments to the Copyright Act (which added, for the first time, an exception for libraries and museums), the Copyright Committee of the national professional association developed a curriculum for a two-day workshop on Canadian copyright law and its application to archives to be available to any association or institution wishing to sponsor the workshop.

2 Half of the questionnaire respondents have a masters degree, although the study did not ask respondents to indicate the discipline.

3 Particularly the national listserv, although many of the provincial archives associations also have listservs for their members.
Just as archivists learn about copyright from a variety of sources, they also consult a range of sources in seeking answers to specific questions. This issue was explored in the questionnaire and in interviews. To respond to the question “Which of the following do you consult if you have a specific question about a copyright matter?” questionnaire respondents were given a list of possible sources and were asked to check all that applied. Questionnaire respondents consider colleagues in the profession, the statute and regulations, and books and newsletters to be the three most important sources consulted in seeking the answer to a specific question. Another cluster of sources (listservs, repository websites, and in-house legal counsel) are consulted by approximately one-third of respondents. Of the ten respondents who checked “other,” five mentioned searching on the Internet. When asked what source they would consult first, archival colleagues (in their own repository or in the wider profession) ranked first (34 respondents), followed by the statute (25 respondents), printed sources (18 respondents), and the in-house copyright specialist or legal counsel (15 respondents).

It is worth noting that nearly half of the questionnaire respondents report that their repository has a designated staff member to be responsible for copyright matters. However, the repository copyright designates do not stand out from their colleagues; differences between the two groups’ questionnaire responses to the questions about their sources of copyright knowledge, the sources they would consult if they had a specific copyright question, and the sources they would consult first if they had a specific copyright question are not statistically significant, suggesting that the copyright designates do not receive any special preparation for their role.

Who or what archivists would consult in seeking the answer to a specific copyright question was explored in more detail in the interviews; the responses further illuminated certain aspects of the questionnaire responses. First, although one would expect that the statute and regulations would be consulted at some point, it is somewhat surprising to find that this source ranks ahead of books and newsletters, given the number of study participants who commented on the complexity and difficulty of copyright law. However, interview data revealed that many archivists are accustomed to consulting legislation in a number of areas that affect their work, and thus are comfortable looking at the copyright statute as a starting point before consulting further sources. Secondly, interview data suggest that seeking an answer to a specific question may require consultation with several sources until an answer is found. Finally, in repositories where archivists have access to in-house legal counsel, the staff often find that the lawyers know little about the operation and professional practices of archives, and they may not be well-versed in copyright law.

Training sessions, publications, and notices about amendments are primarily a means of learning about the provisions of the Act, but they may not address such matters as crafting a copyright policy, or the terms and conditions that a repository may put on further uses of digitized holdings made available on website. For this aspect of copyright knowledge, there is evidence on the websites and in the interview data that repositories base their institutional practice on procedures, forms, and policies borrowed from other repositories, particularly from their websites.

To sum up, archivists’ sources of copyright knowledge are diverse, ranging from structured training events to publications to professional colleagues. It is logical to assume that what Canadian archivists know is in large part determined by their sources of knowledge about copyright, and, if the sources of knowledge are accurate and current, it is likely that the quality of archivists’ knowledge will reflect that. However, not all of the written sources from which archivists obtain their copyright knowledge are authoritative or up-to-date. For example, the printed reference source mentioned most frequently was published in 1999, and has not been updated and reissued. Professional colleagues are reported to be an important source of copyright information, but they themselves may not always be well-informed.

4. QUALITY OF ARCHIVISTS’ COPYRIGHT KNOWLEDGE

The second research question (What do Canadian archivists know about copyright?) was intended to discern the quality of Canadian archivists’ knowledge of copyright. For the purposes of this study, quality of knowledge was understood to reflect accuracy and currency, i.e., the extent to which Canadian archivists correctly understand the relevant provisions of the Act, and the extent to which their knowledge is up-to-date in that it reflects recent amendments and case law.

The study was not designed to administer a comprehensive “test” of archivists’ copyright knowledge, and the findings must therefore be used cautiously. Nonetheless, within the limits of the study, it appears that Canadian archivists’ knowledge of copyright is uneven, both in terms of accuracy and currency. While some archivists appear to have a correct understanding of particular provisions of the Act, others misunderstand aspects of copyright. Even when dealing with a straightforward provision such as the date when photographs enter the public domain (and thus may be digitized and made available online without having to obtain the permission of the rights holder), more than half the repositories that provide data about this issue are inaccurate in some way. Some misunderstandings have few serious consequences, but others may reduce online access to holdings, increase access (but put the repository in a position of infringing copyright), or mislead users (and possibly other repositories, given that repositories borrow procedures, forms and policies from one another). As far as the currency of archivists’ knowledge is concerned, interview data suggest that archivists’ awareness of recent changes to the Act is lacking altogether, or is confused and incomplete.

Incorrect or out of date copyright information may affect access in terms of selection; for example, if a repository’s policy is to select for digitization only material in which the copyright has expired, a mistaken understanding of the provisions for the duration of copyright could mean that material in the public domain would not be selected for online access. Repository websites also frequently contain information about how the online content can (or cannot) be used; a mistaken understanding of copyright may limit legitimate uses, and, where that information is publicly available on websites, it may mislead users and other repositories. In other words, practices based on incorrect information may compromise archival repositories’ mission to make their holdings available for teaching, learning, research, or pleasure.
5. ARCHIVISTS’ PERCEPTIONS OF COPYRIGHT

Repository copyright practices will also be affected by how archivists perceive copyright, i.e., if they see copyright as a problem (for whatever reason), they are more likely to be conservative in their approach to the selection and presentation of holdings they make available online. Questionnaire and interview data revealed something of archivists’ perceptions of copyright. More than half (52%) of 104 questionnaire respondents disagreed or strongly disagreed with the statement, “The risk of legal consequences for copyright infringement involving archival material is low.” Thirty-seven percent agreed or strongly agreed, i.e., they think that the risk is low; 11% were neutral. In other words, a bare majority of respondents believe that there is some risk of legal consequences. However, 80% of 106 respondents disagreed or strongly disagreed with a second statement, “Copyright is not a problem for archival repositories making archival material available on the Internet.” Only 14% agreed or strongly agreed with the statement; 6% were neutral. In other words, most questionnaire respondents believe that copyright is a problem for archival repositories making archival material available on the Internet. In sum, while 80% believe that copyright is a problem, fewer (just over half) believe that there is some risk of legal consequences. The responses to both statements, taken together, suggest that Canadian archivists think that copyright is problematic.

Interviewees, when asked what they thought of copyright as it affected their jobs, considered copyright to be part of an archivist’s job, despite the fact that some commented on negative aspects such as its complexity or the time it took. Six interviewees described it in strong negative terms as “an impediment,” “a nightmare,” “restrictive,” “very tricky,” something that “complicates the job,” or something they wish would go away. Questionnaire responses to an open-ended question that allowed respondents to comment on any aspect of the questionnaire topics revealed similar views that suggest that archivists are somewhat apprehensive about copyright because they lack the knowledge, the time, and the tools to feel confident that they are not infringing copyright in digitizing their holdings for online access.

6. DISCUSSION AND RECOMMENDATIONS

It appears that Canadian archivists go to some effort to learn about copyright in order to comply with the law, but they find copyright complicated and difficult. The study points out some problems in the ways that archivists learn about copyright. Professionals such as archivists, who are not lawyers but are responsible for compliance with copyright (and other laws), need tools and resources to assist them in better managing their copyright responsibilities. This study suggests that the training offered has not been fully effective, and there are a number of weaknesses in the completeness, currency, and availability of the resources available to archivists and their repositories.

Such matters bear further investigation. Replicating the study in other jurisdictions or in related domains (i.e., libraries and museums) may reveal alternative means of delivering copyright education that may be more successful. It would also be interesting to do what this study did not do, i.e., investigate exactly what archivists do know about copyright. The study’s findings suggest that copyright education is very much a matter of continuing education undertaken by professional associations after people have entered the profession; another fruitful area to investigate is the extent to which copyright matters are covered in graduate education.4

While these matters merit further investigation, this study points to an area that needs more immediate attention. Archivists are apprehensive about the copyright issues involved in digitizing their holdings to make them available on the Internet because they lack the knowledge and tools to deal confidently with copyright. Therefore, we must look at ways of addressing those gaps. The following recommendations point to possible courses of action.

The study participants relied heavily on post-appointment continuing education offerings to learn about copyright; however, this suggests that cultural heritage practitioners should learn about copyright much earlier, i.e., as part of their graduate education. While no systematic study has yet looked at the extent to which copyright is taught in LIS programs in North America, a preliminary review of information available on websites of Canadian library schools suggests that, while copyright is presumably covered (to some degree) in courses on reference services, information policy, digital libraries, and the like, it is not taught in depth. Thus the first recommendation would be to provide robust copyright content in graduate programs. Such content must cover not just the provisions of the law, but also how the law operates in relation to the mission of cultural heritage institutions, and the role of the archivist/librarian/curator in administering copyright. Work has begun in this area: the collaborative work underway at Virginia Tech and the University of North Carolina to develop a curriculum for digital libraries will include a module on intellectual property [7], and the DigCCurr project to develop an international curriculum to support the management and preservation of digital materials across their life cycle will also presumably address copyright issues in a way that recognizes differences between national copyright regimes.

Digital and communications technologies have profoundly challenged copyright law; in response, rights holders are seeking clarification through statutory amendments, case law, and new business models and practices. However, the application of copyright in the digital environment is still fraught with uncertainty, and copyright law continues to evolve. Thus, it is essential that cultural heritage professionals keep up with changes to copyright law. One part of this is the need to make practitioners aware of changes and their implications for practice in a timely fashion. This is often done by a committee or working group affiliated with a professional association; the members of such committees are usually volunteers with an interest in the topic, and they notify the membership of the changes through listserv postings. The study suggests that such a model is effective.

Alerting their members to amendments or related case law is important; however, systematic communication about changes must be supplemented by robust continuing education programs. Programs whose content is timely, accessible (i.e., at a reasonable cost in terms of fees, time, and pre-requisite knowledge), and authoritative, are essential to keep stewards of cultural heritage materials informed of relevant changes and their implications for

4 The author is currently engaged in such a study.
practice. Such continuing education offerings are usually delivered by practitioner-contractors on behalf of professional associations or the continuing education units of graduate programs; their quality and timeliness are dependent on the availability of qualified instructors. In order to be timely and effective, the sponsors of such offerings must be committed to work closely with those individuals who have the time and expertise to keep abreast of changes, and be prepared to organize and deliver a suitable training package in response to changes.

The study also revealed that many archivists want to seek solutions to copyright problems themselves, but they lack authoritative, up-to-date, readable reference materials that go beyond the provisions of the law. Practitioners want an interpretive guide to assist them in determining how to proceed when faced with particular copyright situations. Compared with print sources, websites are readily accessible and can be quickly updated; however, someone has to be responsible for the production and maintenance of such a guide to ensure that changes are reflected in a timely fashion. Whether this should be the responsibility of professional associations or some other body, and the extent to which that entity would be prepared to go beyond generalities to avoid legal liability, are difficult questions to which there is no clear answer.

Closely related to this is a need for access to a copyright expert, i.e., someone a practitioner could contact with a specific question and get an authoritative answer, instead of making the rounds of possibly ill-informed colleagues. While there are individuals who possess the necessary expertise (i.e., knowledge of the law combined with a thorough understanding of the mission and operation of cultural heritage institutions), they would presumably expect to be fairly compensated for their time and expertise, and it is not clear who would pay for such a copyright consultation service.

Copyright is an important professional matter for those who preserve and make available cultural heritage material. However, copyright is often complex, time-consuming, and somewhat intimidating; furthermore, it is just one of many aspects of their work competing for attention and resources. Consequently, practitioners do not always address copyright issues correctly, due to a lack of time, in-house expertise, or the financial resources to engage outside expertise. This paper suggests possible ways of addressing weaknesses in how practitioners learn about copyright and keep their knowledge current. A greater level of copyright expertise on the part of practitioners will in turn increase their confidence in their knowledge and their ability to take a less conservative approach to copyright, resulting, perhaps, in more cultural heritage resources being available online.

7. REFERENCES


Building Australia’s eResearch Capability: The Challenge of Data Management

Adrian Burton
Australian National Data Service
Australian National University
Canberra, ACT, 0200, Australia
adrian.burton@anu.edu.au

Margaret Henty
Australian National Data Service
Australian National University
Canberra, ACT, 0200, Australia
margaret.henty@anu.edu.au

ABSTRACT
The creation of the Australian National Data Service (ANDS) provides an opportunity to devise a national approach to the provision of skills for improving both institutional and individual eResearch capability. One of the four proposed ANDS programs is called Building Capabilities. This will have three broad areas of activity: including curriculum development, stakeholder engagement and the development and implementation of an audit and certification framework. This paper discusses the first of these: its target groups and constituency, skills areas and levels, and delivery strategy.

Keywords
Capability building, eResearch, learning, skills development, maturity models

1. INTRODUCTION
The importance of data in the scholarly communications cycle is becoming increasingly apparent. More data is being created, usually now in digital form, and our capacity to process and transform it has grown. Data is seen increasingly as critical to the resolution of major world challenges such as climate change and energy. Recognition of the need to improve our use and management of data has led to the creation in Australia of the Australian National Data Service [1].

ANDS will move from its planning phase to functional reality in the latter half of 2008. There will be four programs of activity designed to help Australian researchers in this emerging environment and to start building the Australian Research Data Commons. They will deal with issues of research data ownership, roles and responsibilities, access, co-ordination and curation. The four programs are Developing Frameworks, Providing Utilities, Seeding the Commons and Building Capabilities.

2. BUILDING CAPABILITIES
One of the proposed programs of activity is that of Building Capabilities, designed to “improve the level of capability for research data creation and management as well as research access to data (and associated technologies) across Australia” [2].

In order to achieve the three-year outcomes listed in the Interim Business Plan, the ANDS Building Capabilities Program will have three broad areas of activity: the development and delivery of a data management curriculum, engagement with key stakeholders and the development and implementation of an audit and certification framework for staff, services and data facilities. Curriculum development and community engagement have immediate priority.

The Building Capabilities Program will assist organizations to grow their skills and capabilities in data management. From the perspective of the organizations themselves this is part of a broader framework to assess and improve their capabilities and capacity to support advanced ICT-enabled research. The activity of this ANDS program, therefore, sits within a broader e-research capability maturity model. This model provides the context, framework, vocabulary and approach for the development and delivery of a data management curriculum.

3. OUTCOMES
ANDS will provide a structured set of modules for delivery. These will target particular groups within our constituency, cater for different levels of expertise, target particular areas of skills and designed to complement discipline- or institution-specific materials. ANDS will work with partners who can develop and maintain materials, and take responsibility for their delivery within their own institutions. The modules will be organized into a framework which will allow people to combine relevant modules into a certifiable program.

4. THE ANDS CONSTITUENCY
The organizations which make up the ANDS constituency include all Australian universities, publicly-funded research organizations such as CSIRO, GeoScience Australia, the Australian Bureau of Statistics, and the cultural collections sector which includes galleries, libraries, archives and museums. The first two of these groups will be given priority in the first instance.

Within the constituency organizations, there are four groups which are seen as targets for training: research staff; IT staff such as data modelers and informatics specialists; data facility and repository managers; and librarians and archivists. We recognize that different individuals and organizations are at different stages of capability, so it will be important to provide curriculum materials appropriate to those different levels. We also recognize...
that all groups need some skills in common but that each has specific requirements over and above those.

5. SKILLS AND SKILLS LEVELS
The skills areas to be targeted have been identified through surveys and interviews with a wide variety of people associated with the conduct of data-intensive e-research and the stewardship of data. These include: advanced scholarly communications, understanding the legal and regulatory environment, data management policy development and planning, information management, informatics and data analysis. This list is neither exhaustive nor exclusive and will be refined on the basis of the availability of existing materials and expertise in partner organizations.

6. DELIVERY STRATEGY
There are many logistical issues to be worked through in the development of the proposed curriculum and some solutions and innovations should be apparent by the time of the conference. These issues include coping with different existing skills levels and finding suitable delivery mechanisms, given that ANDS funding does not extend to resourcing the delivery of capability building programs to all ANDS constituency groups. Training trainers will be a high priority.

7. ACKNOWLEDGMENTS
ANDS is supported by the Australian Government through the National Collaborative Research Infrastructure Strategy (NCRIS).

8. REFERENCES

The Survival of Records (and Records Management) in the Twenty-First Century

Kenneth Thibodeau
National Archives and Records Administration
Washington DC, 20740 USA
Ken.Thibodeau@nara.gov

ABSTRACT
This paper discusses the nature and role of records and records management in relation to the contemporary cyber-landscape, and describes how the principles of records management inform the retention of electronic records and what changes in perspective and method are needed given the digital domain.

Keywords
Electronic records, recordkeeping, digital preservation, InterPARES.

1. INTRODUCTION
The first thing that might come to mind, on hearing the topic of the survival of records in the twenty-first century, is the problem of digital preservation. Obviously, the preservation of records in digital form is crucial, but while preservation is necessary, it is not sufficient. Given the open-ended proliferation of new forms for creating, capturing, and combining information in digital systems, of ever wider, more diverse, and more highly specified ways of applying digital technology in the conduct of business, of the expanding capacity to re-use and even re-purpose digital data to satisfy a variety of both planned and spontaneous needs, it is also necessary that records continue to be recognized not only as a distinct class of information assets, but also as one which merits special attention. It is not sufficient for records keeping to be seen as a necessary part of doing business. Necessity of this sort is all too often the daughter of laws, regulations, and other external requirements. Records that are kept as the result of such external forces are easily relegated to the sidelines and perceived as having marginal value in the accomplishment of practical objectives, in strong contrast to the immediate value of real time online transactional processing, multidimensional analytic processing, geo-referencing, experiential computing, and other powerful tools made available by the application of computer science and engineering. If records slide to the margins of the conduct of business, records management will diminish with them. To survive and prosper in the twenty-first century, records and methods for managing them need to have a vibrant, organic relationship to the conduct of affairs. To be a vital contributor to corporate, institutional, individual and societal success in the twenty-first century, records management must deal cogently and comprehensively with the increasing permeation of digital information in the conduct of affairs.

Cyberspace is so different from what has gone before that it has been characterized as the fifth dimension, one which enables us to create and do things which are not possible in the space-time continuum. The creation, capture, and communication of information in cyberspace is drastically different than in four dimensions, because the digital dimension breaks the bounds of space and time that constrain the information technology of hard copy documents. To face the challenges and take advantage of the possibilities offered by digital technology, we need a richer and deeper understanding of the nature of digitally encoded information and how such information can be and can be managed as records. We need to be better able to apply the knowledge we have of records and records management in the digital realm; translating it in terms that make it effectively operative in cyberspace; adapting it where necessary; and also abandoning those concepts and techniques that are not viable in cyberspace. If we fail to do so, we run the risk of seeing records management become an increasingly esoteric exercise, divorced and isolated from the mainstream of affairs.

The first thing records managers need to do is to acknowledge that established knowledge and methods have limited applicability in the digital dimension. The second thing – logically but not necessarily chronologically – is to recognize the opportunities that digital technology creates for managing records in ways that might exceed by far anything possible in four dimensions. Consider just two basic topics: What are electronic records, and how should we organize them? For both topics, let us consider how well or how far established knowledge applies in the digital dimension. Some obvious limits to its applicability, and alternatives that the digital dimension opens up.

2. WHAT ARE ELECTRONIC RECORDS
Conceptually, electronic records are not radically different than traditional records. Traditional definitions and concepts identify fundamental characteristics that are independent of how records are constructed, encoded or stored. A record:

- is a unit of information;
- is made or acquired in the course of activity;

1 Christopher Greer. 2007.
• contains evidence or information about that activity;
• is kept for use in subsequent activity or for reference, and
• is organically related to other records of the same
records creator and activity.²

But problems arise when we try to deal with electronic records empirically. From an empirical perspective, we are not yet able to characterize them in a way that would enable us to articulate surefire methods and processes to maximize their value in the conduct of affairs. Records as we know them are artifacts of a particular genre of information technology, the technology of hard copy, where information is affixed to a physical medium in a hard and fast way. Digital encoding fundamentally alters the relationship between medium and message. In the digital realm for all practical purposes the relationship between medium and message is immaterial. Digital data move frequently and repeatedly from one physical medium to another: from silicon chips to magnetic drives, solid state memory, optical discs, copper cables, optical fibers, and electromagnetic waves, but the information objects the data constitute persist unchanged across such physical transformations.

The message here is not the medium, and it is certainly not about the durability of the medium. In the realm of hard copy, people could, did, and do use both permanent and short-lived media: from the clay tablets of the Babylonians to the wax tablets of the Romans, from cheap newsprint to archival quality paper. The same is true in the digital realm. There is no physical or chemical barrier to permanent digital media. You could write digital data on clay tablets. It’s just that be business case for doing so is lousy.

The basic difference in the relationship of electronic records to their physical carriers is but one of many ways that electronic records substantially differ from traditional records. Additional differences will emerge as digital technology evolves. Mennen-Haritz suggests that the shift from ‘written’ to electronic communications is as epochal as the shift from oral to written communications in past millennia. Articulating a variation on the theme that form follows function, Mennen-Haritz points out that changing forms enables and impels changes in functions and that, conversely imposing old forms may constrain our ability to function effectively [1]. New biological species emerge through the gradual accumulation of mutations operating on a very small scale, but widespread speciation, as well as its evil twin, extinction, is often driven by large scale disruptions, such as global changes in environmental conditions. We should expect that in the twenty-first century some older forms of records and ways of managing them will not survive. They will either become extinct or retreat into ecological niches in the information landscape. We should also recognize that both marginally different and radically new types of records have already emerged in cyberspace, including new genres that have no parallel in the world of paper and other hard-copy records. With types of electronic records which appear to be counterparts of familiar hard-copy documents, we must be able to recognize and respond to even minor mutations that either imperil or promise to improve the value of records. New species of records will continue to emerge apace with the evolution of digital technology.

A basic difficulty we encounter in trying to apply knowledge derived from experience with traditional records in cyberspace is how to identify an individual record, an item of information that cannot be further decomposed without loss of ‘recordness.’ In physical space, unit records are often congruent with the physical media on which they are inscribed: a piece of paper, or several pieces stapled or clipped together, or a roll of motion picture film. In cyberspace, what appears as a single document may actually consist of data that are stored in numerous separate objects, each with a different structure and semantics; assembled by means of an intermediate object, such as a view on a database; organized according to the specifications of a form; and presented according to the dictates of one or more style sheets. There is nothing – no single object – stored in the computer system that corresponds to the document presented to a human in such a case. Such a document – which I call a “pseudo-document” or “pseudoc” for short – fails to satisfy one of the defining characteristics of a record: it is not kept.

How then do we deal with the case where a pseudoc is the exact counterpart of a traditional record and serves the purposes the traditional record would have served? The first InterPARES project studied numerous cases of this sort and concluded that it is literally not possible to preserve such records in electronic form. It is only possible to preserve the ability to produce copies of the records. This led to the articulation of sets of benchmark and baseline standards for maintaining the ability to produce authentic copies of such records [2]. But, if we are not preserving the records, but only the ability to reproduce them, what is the stuff we do keep in digital form? Analyzing cases which encompassed both digital replicas of hard copy records and electronic records which have no traditional counterparts, Duranti and I arrived at the necessity of distinguishing two different classes of records in such cases: the information kept in digital form, which we designated as a ‘stored record’ and the rendering of this information in a form suitable for human use, characterized as a ‘manifested record’ [3].

Keeping a manifested record in the manifested form would be redundant given the stored record. Thus, the term, ‘manifested record,’ is shorthand for “a copy of the record we would have kept had we decided to keep it in human readable, rather than digital, form.” The manifested record may exactly reproduce the content and form that a human author saw when the record was created, but it may also be a document that a computer application produces de novo from data extracted or derived from the contents of one or more other documents created either by human authors, by processing of externally originated data, or by system to system interactions. The output from one or more stored records is not necessarily a record. It may be a temporary, evanescent display which is not saved. Nevertheless, given proper procedures and controls, a manifested record can be an authentic copy of a record.

How do we identify and delineate individual digitally stored records? In simple cases there can be a one-to-one correspondence between a digital file and a manifested record. But in many cases, the relationship between stored and manifested records can be many-to-one, one-to-many, and even

² See, for example the definitions in the ISO Records Management Standard, ARMA, the InterPARES project.
many-to-many. The diversity and complexity of ways that pieces of information can be stored digitally and combined and processed either to generate a human-readable manifestation or for uses in system-to-system transactions led Batley to argue that in cyberspace we should concentrate on “information chunks.” Psychology defines an information chunk as a unit of perception and meaning, or a digestible unit of information. Batley prefers the definition of information chunks within documents offered by information architecture gurus, Rosenfeld and Morville: “The most finely grained portion of content that merits or requires individual treatment” [4]. Batley asserts that “Each chunk will have a status and a function (and a form) that will merit its definition as a discrete record. Clearly then, it is not the document itself that merits the status of a record, rather it is the chunks of information content each document contains that are the records that need to be managed” [5].

This approach, however, fails another of the defining characteristics of a record: its relationship to an action or activity. Some information chunks can be used in many records of many different actions. For example, data identifying a customer will appear in every interaction with that customer. They will appear even in multiple records related to a single transaction, such as the order, shipping manifest, and invoice for a sale of merchandise. While the data identifying the customer are necessary parts of the contents of such records, the chunk itself carries no clue of the actions or records in which it participates. Batley's argument that managing electronic records entails managing information chunks aligns with a way computers commonly identify, store and process information, but the concept of raw information chunks needs to be supplemented in order to provide even a minimal basis for either providing adequate information about past activity or contributing to subsequent activity. The InterPARES project distinguished information chunks that occur in or contribute to the reproduction of manifested records, calling them “digital components” of electronic records. A digital component is a bit string that is necessary to produce a manifested record. A digital component might be called a “record chunk,” but it is below the level of a record. To rise to the level of a record, an information object must have a defined, or at least definable, relationship to activity. We need to manage at the level of digital components, as well as at the level of records, in order to ensure that we can reproduce manifested records from their digital components.

Digital components should not be limited to portions of content, as they are in the Rosenfeld/Morville definition. A digital component might not consist of content, but define what content should be included in a manifested record, such as in a database view. It could also specify the semantics, syntax, or presentation of the record; for example, a statement in an XML Schema, an Xpath query, or a Cascading Style Sheet. Basic objects, such as the dynamic load library for a face or a color space should be treated as digital components if they are necessary for the output of an authentic copy of a record. Overall, there are four categories of digital components of records: (1) composition data, which tell the system what form and content data belong to a document, (2) the content data, (3) the form data, which determine how the content is arranged and presented, and (4) rules. Several different types of rules can shape the production of a manifested record. For example, rules may define the conditions or circumstances in which the record can be reproduced, or they may exclude or include certain elements of content, depending on a user's privileges, or they may define links or hyperlinks between documents or parts of a document [3].

Again, they should be retained as long as any manifested record which they control is needed and they should be managed to ensure that authentic copies can be produced.

But we are still left with the question of how to distinguish individual stored electronic records in cases where there is not a one-to-one correspondence between a stored and a manifested record. A record is something that is kept. Therefore, a stored electronic record must be a persistent object that is maintained in a computer system. A record provides information or evidence about an activity or a state of affairs at a particular time; therefore, the persistent object must contain fixed, invariant data. But it is not necessarily the case that the stored record itself provides complete information about one or more actions or a particular state of affairs. This is not peculiar to the digital realm. At least since World War II, it has been extremely common that the complete 'record' of a single action is contained in many different records. Furthermore, a stored electronic record does not necessarily provide such information or evidence directly in the form in which it is stored. There is no a priori reason why an actor could not create records that require some combination with other records, or some processing in order to deliver meaningful information about past actions or situations. Again, this is not peculiar to cyberspace. Many governments, for example, do not keep birth certificates as distinct documents. Rather they capture the necessary data about each birth in a registry that contains the same data elements about all births within the jurisdiction. An individual's birth record can be produced on demand as a separate document by copying the relevant data from the registry onto a blank birth certificate. This is essentially analogous to the production of a manifested record from a stored digital record, with notable differences that computers provide much more flexibility in how the data about individual activities or situations are recorded and – because they not only capture, organize, and store data but also participate in the execution of business processes – can more reliably enforce data quality by embedding business rules in the execution of processes. Thus, a stored electronic record can satisfy the requirement that records provide information or evidence if it can be used to produce one or more manifested records that directly communicate such information or evidence.

We can, then, formulate a simple set of three criteria for identifying a stored electronic record: it must be (1) a persistent digital object that (2) contains fixed information about an activity or the state of affairs at the time when the action was done, and that (3) can be used to produce one or more manifested records. The stored record may be a single information chunk or digital component, but it might just as well contain many thousands of such elements. There is no a priori limit to the structure or content of a stored electronic record and it may contain data in any one or more of the four types of digital components; that is, content, composition, form, or rule.

3 Gerard Naud and Christine Naud. Gazette des Archives.
4 The third criterion should be expanded to include records which are not manifested, but produced from stored records for interactions between systems where no human is involved.
3. HOW SHOULD WE ORGANIZE ELECTRONIC RECORDS

The properties of records inevitably impact methods for managing them. Methods that we have come to think of as fundamental to managing records can be seen on reflection also to be artifacts of hard copy technology. Traditional filing systems are based on the physicality of hard copy records, on location or more specifically on collocation in file folders, filing cabinets and file stations. It is certainly possible to import such approaches into cyberspace, at least iconically, as we see in Windows file management and email management products. This approach is embodied in records management applications that implement the Department of Defense standard, DoD 5015.2-STD. These applications effectively create virtual filing cabinets in cyberspace and allow us to manage electronic records as if they were in metal cabinets. But possible does not equate to optimal. What is important in the digital realm is neither the media on which information is recorded, nor the physical place where it is stored, but the possibilities that digital technology creates both for the forms of information that can be created and the ways they can be organized and used. The diversity of ways in which data can be organized; for example, in relational and object-oriented databases, data warehouses, and geographic information systems, are obvious advantages of digital technology, as are the possibilities for multiple simultaneous arrangements and for virtually instantaneous recombinations.

Do records exist in such applications? Can they exist? We know that it is certainly possible to produce manifested records from such applications. Selected content of databases is commonly output in the equivalents of traditional forms and reports. But the output of a system is not the same as what it contains. If so, how can we manage the records? One possibility might be to take them out of such applications and put them into virtual filing systems. But to do would diminish their usefulness in the conduct of business for the basic reason that they would exist apart from the systems used to execute business processes and most likely in formats that would not be useful in these processing. Breaking the organic links between records and activities would also diminish their value as records. If you want to know what an organization did and you had two places to find: one the system the organization used to conduct its business and the other a repository where it put special forms of information that satisfy some abstract criteria for evidence, which would be the better source? Other things being equal, the system that contained the information used in the conduct of affairs, in the forms in which it was used, would be the better source. But isn't the fundamental purpose of creating and keeping records to provide a privileged source of information about prior activity? Does this represent a dilemma? How can we manage records if we don't put them in filing systems? There are other ways, made possible by digital technology. One method that is emerging is called Records Management Services. It was initiated by the National Archives and Records Administration, articulated in collaboration with 19 other federal agencies and is now being developed, with much broader participation, as a voluntary standard under the aegis of the Object Management Group. In brief, Records Management Services provide methods for identifying records in practically any type of computer application and form managing them within their native applications [6].

4. THE ROAD TO PROSPERITY AND SURVIVAL IN CYBERSPACE

Can we discern a path for records management that will enable it to survive and prosper in cyberspace? Certainly, the basic role of helping an organization to determine what records it needs to create and keep remains essential in cyberspace, with some adaptation, such as determining what records the creator needs to be able to manifest, rather than to keep. While records managers cannot personally have sufficient IT expertise even to implement decisions about what records to keep, they must be able to work with and guide a range of IT specialists in developing and maintaining systems that meet requirements for records creation and retention. But activities like this essentially translate traditional activities into the digital realm. That is not likely to be sufficient for records management to truly prosper.

Records management can contribute to realizing the potential value of digital technology in a way that best satisfies the needs of organizations and individuals in the conduct of their affairs. To do so it must be able to identify and show how records exist in business systems and it must offer methods for managing electronic records that convincingly add value in the conduct of affairs. For this, the discipline must move beyond established knowledge and methods.

Not only the specific knowledge, but even the kinds of knowledge that have stood us in good stead in the hard copy world cannot migrate into cyberspace. We should not seek to develop the digital equivalents to the kinds of expertise we had, for example, in the relative merits and drawbacks of end digit filing, or the differences between diazo and silver halide films. For one thing, the digital realm is too big and complex for us to develop sufficient knowledge about the different ways information can be encoded, organized and used. For another, by the time anyone could develop such mastery of any particular digital data type, the technology would have changed.

Information and networking technologies have substantially raised the bar on expectations for information used in the conduct of affairs, and they will raise it even higher in the future. Both competition in the market place and improvements in computer science and engineering expand the range and the amount of things digital technology can do for us, increase the flexibility and speed with which they can be done, reduce associated costs, and even make it possible to do things that are impossible outside of cyberspace. Hallmarks of raised expectations include availability, ubiquity, fluidity and responsiveness. Availability: the information needed at any time should be easy to find and access. Google is the norm and the browser is increasingly the form. Ubiquity: given the Internet, intranets, wireless, and the convergence of voice and data communications, information should be available wherever it is needed, regardless of where it is stored. Fluidity: it should be easy to change both the content and the form in which information is presented in response to different needs. This includes both adapting data to the device on which it is delivered, from a desktop monitor to a hand held PDA to a telephone, and the ability of users to select among different data types; such as, “Click to enlarge” images and tabs to select either product specifications or customer reviews in an online sales application. Responsiveness: information should be suited to business needs regardless of how complex, convoluted, demanding or dynamic they are. We already rely on digital data
to build, operate, and maintain aircraft carriers; to identify specific genes that make certain individuals more susceptible to particular diseases; to find and guide us to a good restaurant in a city we are visiting for the first time; and not only to conduct wind tunnel experiments, but even to modify those experiments in real time based on the performance of different sensors. The range, richness and usefulness of digital technologies are likely to increase.

What, then, is the potential for records management to add substantial and readily recognized value in this context? Just as the basic definition of ‘record’ does not need to change in cyberspace, the basic knowledge that records managers have of the most important information assets an organization has, its records, is a solid foundation for helping the organization transition to and operate in cyberspace. After all, records managers are not alone as neophytes in the digital dimension. Indeed, records managers are advantageously positioned to help organizations navigate in cyberspace because of the simultaneously broad and deep insight they have into what types of information are used in what parts of an organization and for what purposes. They can use this advantage to promote one of the basic objectives of records management: to ensure that all of the right information and only the right information is delivered to the right person at the right time to meet organizational needs.

Optimizing the utility of information entails the superposition of three architectures: information architecture, system architecture, and social architecture [7]. Information architecture is the organization of an information space, including the information it contains, to facilitate discovery and delivery of required information. System architecture is the organization of information technology components to provide functional, data and other support for specified requirements. Social architecture is the organization of social networking and other social media tools and user experience capabilities to optimize the usefulness and use of digital information. Records management knowledge is most closely aligned with information architecture. It is not of much use in building or sustaining systems architecture, and effectiveness in using social architecture tools requires a very different type of expertise. However, records managers can leverage their overarching knowledge of what types of information are used in different parts and processes of an organization to contribute to a coherent approach to optimizing information assets, one that spans all three architectural domains.

5. REFERENCES
ABSTRACT
This paper discusses the impact of digital publishing, e-mail, and electronic records management on the North Carolina State Archives and the State Library of North Carolina, the entities responsible for gathering, providing access to, and permanently storing state agency electronic publications and records in North Carolina. In addition to outlining the current state of digital government information in North Carolina, as evidenced by recent survey data, this article touches on future plans and collaborative efforts, both within state government and with other states, as well as some of the challenges to successful implementation that the State Archives and State Library must overcome.

Categories and Subject Descriptors
H.3.6 [Information Storage and Retrieval]: Library Automation—Large text archives; I.7.4 [Document and Text Processing]: Electronic Publishing; H.2.8 [Database Management]: Database Applications—Spatial databases and GIS; H.4.3 [Information Systems Applications]: Communications Applications—Electronic mail; H.3.7 [Information Storage and Retrieval]: Digital Libraries; H.3.5 [Information Storage and Retrieval]: Online Information Services—Data sharing; H.3.5 [Information Storage and Retrieval]: Digital Libraries; H.3.5 [Information Storage and Retrieval]: Online Information Services—Data sharing; K.5.2 [Legal Aspects of Computing]: Governmental Issues—Regulation; K.6.4 [Management of Computing and Information Systems]: System Management—Centralization/decentralization.

Keywords
Digital curation, state government information, collaboration, e-mail, digital publishing, geospatial data, state publications, geographic information systems, digital preservation, electronic records.

1. INTRODUCTION
The State Library of North Carolina and the North Carolina State Archives are the official repositories for publications and records, regardless of format, produced by all state government agencies. For years, the acquisition, management, access, and preservation of paper publications and records has been systematically approached — albeit not without challenges and shortcomings — through such programs as the State Publications Clearinghouse of the State Library [1] and the Government Records Management Program of the State Archives [2]. Like many states, North Carolina has a decentralized government structure, but in the world of paper, publications were typically produced and distributed through the office of the agency chief information officer (CIO), making the acquisition by the State Library somewhat less challenging than contacting every author at every agency and division throughout the state. Likewise, records were transferred to the State Archives according to records retention schedules through an established workflow and authorization process.

With the coming of digital publishing, e-mail, and electronic record keeping, the world changed dramatically for producers of information dispersed throughout North Carolina state government, the CIOs, and the State Library and State Archives programs responsible for their curation. An already decentralized government became even more so as publications went from formal documents with formal publishing policies to digital files that could be produced quickly, with little intervention from designers, editors, or agency officials. Paper records that were once relatively stable and "easy" to store became electronic files that could be modified, were poorly named or indexed, or disappeared as quickly as they were created. To this day, if and where many of these files reside remains a mystery. For the State Library and State Archives, the divisions of the N.C. Department of Cultural Resources (DCR) with the legally mandated responsibility for the long-term maintenance of publications and records, and for the citizens of North Carolina, who demand an open and transparent democracy, this is not just a problem to ponder, but one that requires an absolute answer.¹

1.1 Publications
In 2003 and again in 2008, the State Library surveyed state agency staff to attempt to identify what was being produced, how, and where electronic files were being stored. The goal of the 2003 survey was to establish a need for a program within the State

¹ The issues North Carolina and other states face in addressing this problem are well documented. Several reports offering some background and context for these issues can be found at http://www.ctg.albany.edu/themes/pubs?chapter=records.
Library that would be responsible for investigating the challenges of long-term preservation of agency publications and implementing solutions as they became available [3]. That goal was reached in 2006 through permanent funding of the Digital Information Management Program (DIMP) and the continuation of a Statewide Leadership grant called the Access to State Information Initiative, now in its fourth year. The 2008 survey followed up with questions regarding current publishing practices and continuing participation in the State Publications Clearinghouse program to help inform the DIMP’s current and future plans for a digital repository [4].

The findings of these surveys have been telling. In general, agencies have not begun cooperative efforts (although some are working intra-departmentally) to ensure long-term management of digital files of any sort, and the trend continues towards distribution, rather than centralization, of digital assets. Currently, 95.5% of agencies publish mainly to their website(s), and have used over 20 different formats to do so in the past three years. Forty-two percent of respondents to the 2008 survey stated that publications remain on publicly accessible sites for only a limited period of time. Over 56% of respondents said that the public is not notified when data is moved or removed from, or updated on agency sites. Once removed, data is often stored in locations not accessible by the public, including agency servers (39% of respondents) and personal computers (32%).

The State Library’s answer to the surveys is a centrally managed, digital repository of born-digital state publications. Ingest occurs in two ways: either the “push” system or the “pull” system. Many agencies “push” their digital publications to the State Library (over 73% of respondents are currently complying with the law) through email or other submission methods. Or, the State Library proactively “pulls,” or acquires, publications by harvesting them from Agency websites using Archive-It software, contacting agency staff via telephone or email to request specific publications, or obtaining the publication through other manual methods. However they are acquired, all publications are made accessible via several web interfaces and databases, including the Internet Archive and a CONTENTdm database. Work to provide access to these publications through a single portal is ongoing. Preservation of these files is, likewise, ongoing, however, archival copies are currently monitored and maintained in several locations.2

1.2 Geospatial Data

The world of cartography has changed from the use of a compass, brushes and parchment paper to create a map to the modern Geographic Information System (GIS), reliant on complex relational databases that integrate digital graphical data collected by satellites and GPS units with tabular attribute information such as street addresses, environmental observations or demographic data. As the tools used to create the map have grown more complicated, the challenge of preserving this information about “places” has become much more difficult. Instead of just preserving the paper map, archivists must now be concerned with capturing the digital data sources, databases, processes and decisions used to generate a map in the current digital GIS era. As the government entity with curatorial responsibility for these files and through participation in two federally funded grants, the State Archives has begun to understand how these files are created in the office and how they might transfer and be appropriately managed.3 Through a Library of Congress National Digital Information Infrastructure Preservation Program (NDIIPP) grant called the Geospatial Multistate Archive and Preservation Partnership, GeoMAPP [5], the State Archives has begun to understand the creation and movement of geospatial content on both the local and state agency level in North Carolina. In the summer of 2008, the State Archives surveyed creators of geographic information systems (GIS) data about current creation and data archiving practices throughout the state. The introduction to the survey clearly delineated the difference between routine backups and data archival practices in an attempt to focus the responses on data that are being preserved for long-term use and analysis [6]. Of the respondents to the state agency survey, approximately half stated that they currently capture or retain snapshots of superseded geospatial data for archival or historical purposes. The other half either responded that they did not retain snapshots or were unsure. Slightly less than half of the respondents named records retention schedules and historic mapping as the primary purpose for retaining their data.

The North Carolina Center for Geographic Information and Analysis (CGIA) conducted a similar survey for local and municipal governments in the summer of 2008 [7]. This survey was a follow up to a survey conducted in 2006 as part of another NDIIPP grant, awarded to North Carolina State University Libraries, called the North Carolina Geospatial Data Archiving Project. The local surveys were aimed primarily at documenting current production of GIS data at the county and municipality level. Responses from the 64 participating counties in the second survey confirmed results from the first: almost 50% of local government GIS coordinators are taking archival snapshots of at least one type of geospatial data annually. In addition, approximately 58 counties publish their GIS data to the viewer and acquisition portal developed by the state of North Carolina called “NC OneMap” [8].

2 Other states are also working to provide long-term access to publications using various tools and approaches. For instance, the Arizona State Library and Archives is leading a multi-state effort to build an automated, integrated, curatorial workflow combined with existing LOCKSS (Lots Of Copies Keeps Stuff Safe) technology through the PeDALS (Persistent Digital Archives and Library Systems) project (http://rpm.lib.az.us/pedals/index.asp). Similarly, the Minnesota State Archives is leading another multi-state initiative to develop A Model Technological and Social Architecture for the Preservation of State Government Digital Information (http://www.mnhs.org/preserve/records/legislativerecords/). This initiative will focus on identifying extensible methods to provide enhanced online access to digital materials by drawing on existing tools already in use by participating state partners.

3 For a broader view of the challenges of preserving and accessing GIS data, see http://www.lib.ncsu.edu/ndiipp/, http://www.ngda.org/index.php.
1.3 E-mail

While publications and geospatial data present preservation challenges, e-mail presents its own challenges to long-term management. No less important than their paper counterparts (memos, letters), but so much more complicated, these files are large in nature and have unique dependencies regarding relationships and attachments that place them in context. During the spring of 2008, the status of e-mail as a public record was affirmed in the findings of the E-mail Records Review Panel created by Governor Mike Easley [9]. This panel included members of state agencies, the public, and the press, as well as university and city representatives. The State Archives, on two occasions, has accessioned e-mail per the records retention schedule: from the office of former governor Jim Hunt, and from former Superintendent of Public Instruction Michael Ward. Practices in North Carolina state government agencies, prior to the panel’s findings, were to simply delete e-mail belonging to employees who left or retired. In the case of the Hunt administration, the State Archives received the entire e-mail server at the end of the governor’s term. The State Archives received Dr. Ward’s e-mail on DVDs that contained his e-mail account in his original folder structure that reflected his categorization of information. In both instances, information technology staff in the office burned the files to disc and transferred it to the State Archives. The other records in analog format were transferred at the same time. ⁴

2. COLLABORATIVE DIGITAL CURATION EFFORTS OF THE STATE ARCHIVES AND STATE LIBRARY

As previously stated, the State Library and the State Archives are legally mandated by general statute to manage and preserve different state government information — the Library is responsible for “publications,” while the Archives is responsible for “records.” In the analog world, these information formats were more clearly delineated, and as such, each institutional division established its own programs to collect, process, preserve, and provide access to the information for which it was responsible [10]. Because these curatorial programs did not support the systematic collection, preservation, and access to state information in digital formats, valuable state information was at risk of being lost. In recent years, especially as the transition to digital creation has accelerated and the lines between the information formats has blurred (e.g., is a website a publication or a record?), the two divisions have realized the value and necessity of pooling resources, staffing, and funding to initiate collaborative curatorial efforts focused on digital content. As a result, the State Archives and the State Library now collaborate not only with one another, but also with other North Carolina state agencies and similar institutions in other states to identify curatorial solutions that will work across state government.

2.1 Collecting and Processing

2.1.1 Websites

One area in which the State Archives and the State Library have an established record of collaboration is in the harvesting of state agency websites. Retention and preservation of State Agency web sites is the responsibility of the Department of Cultural Resources. ⁵ In 2002, the State Archives did a test of a “manual” transfer of its own website to be accessioned into the Archives. The process became tedious, time-consuming, and was not accurate and true to the information it collected. The webmaster included all the files on the web site in addition to any files he may have created and not used on the site. In 2005, during a pilot testing of new technology, it became clear that centralized web harvesting through the State Library and State Archives was more economical, efficient, and effective than asking each agency to identify its own solution. Starting in 2005 with the Internet Archive’s Archive-It [11] harvesting service pilot, North Carolina has archived over 25 million files comprising approximately 2.5 terabytes of storage in the form of web pages, publications, reports, videos, and other important state government information that might otherwise be permanently lost. Former Governor Jim Hunt was the first Governor in North Carolina to have a website. Although repeated attempts were made to obtain a copy of the site, the State Archives never received one. However, during the Archive-it pilot, the State Archives and State Library captured the site. Soon after, the site was taken down from the live web, and now the Web Site Archives is the only instance of this historic web site. The State Library’s State Publications Clearinghouse has taken advantage of this service to locate copies of many electronic state publications not submitted through the standard “push” methods. Likewise, the State Archives is able to ensure that regular snapshots of electronic records posted on the web are captured. In addition, this single solution for the two divisions has produced a single search interface across all harvested North Carolina state government web content.

2.1.2 Publications

The State Library currently employs a combination of hosted vendor tools that together answer several, specific needs. These systems are neither interoperable (although work on at least some of them is promising in this area) nor does the entire set answer all of the Library’s current digital curation needs. Consequently, the State Library will collaborate with the Washington State Archives as part of the Multi-State Preservation Consortium to test the functionality of the Washington State Digital Archive with respect to electronic publications and certain record series [13]. Testing this Microsoft products-based system is of great interest to North Carolina because the NC Office of Information Technology Services (ITS) has existing purchasing contracts with Microsoft, making replication of the Washington State system a greater possibility. The hope is that the Washington State Digital Archives tool will offer a seamless, end-to-end digital information management solution (acquisition/ingest, management, access, preservation) that meets the needs of both the State Library and


⁵ For citation of the retention schedule item for Web sites, please see http://www.records.ncdcr.gov/schedules/GS_Amendments2006.pdf.
the State Archives so that resources can be pooled to jointly manage the system in house.

Under the auspices of a National Historical Publications and Records Commission (NHPRC) grant to the Data Intensive Cyber Environments (DICE) group, the State Archives and the State Library are also working with local universities via the North Carolina-based Renaissance Computing Center (RENCI) to evaluate iRODS™ (Integrated Rule Oriented Data Systems) as another potential preservation environment [14]. Currently, iRODS does not offer a preservation mechanism that will allow for the rendering of data over the “long-term,” but the developers are looking at incorporating one of two preservation mechanisms in testing. And because iRODS is infrastructure independent, it can be used in conjunction with any of the other solutions the State Library and State Archives are investigating [15].

2.1.3 Geospatial Data
The State Archives does not currently collect or accession geospatial data. One component of the GeoMAPP grant is to determine the feasibility of collecting this data and building the capacity to provide access to it. This is a collaborative effort with CGIA; Kentucky’s Department for Libraries and Archives (KDLA) and Commonwealth Office of Technology’s Division of Geographic Information; and Utah’s Division of Archives and Automated Geographic Reference Center. A second GIS grant based in Raleigh (the North Carolina Geospatial Data Archiving Project at North Carolina State University [16]) has provided the State Archives with greater insight into the skills, knowledge, and resources required to institute an archiving program of this magnitude.

Since there is no single entity designated or processes established for collecting GIS data in North Carolina, it remains dispersed throughout the state. Some attempts have been made to create access points; namely, NC OneMap, the North Carolina GIS portal. NC OneMap provides discovery and access to North Carolina’s geospatial data resources from 70 of North Carolina’s 100 counties. The level of county participation differs: in some instances NC OneMap contains an individual county’s GIS data, while in others, it simply acts as a portal to data located on county servers. GIS data from participating counties can be downloaded via FTP from the NC OneMap web site, or transferred on a portable hard drives upon request.

GIS file formats for distribution, access, and preservation vary from county to county and there are no defined packaging standards. Shape files are the de facto file format for distribution. However, in order for the data to render logically and “live in coordinate space,” five files are required for distribution. If any of the five is missing, the data must be “massaged” or manipulated in order for it to be understood.

2.1.4 E-mail
The State Archives first began accessioning e-mail in 2001 after Governor James Hunt left office. At the end of his term, the State Archives received over 6 gigabytes of e-mail. Much of this had not been culled and State Archives staff had to appraise the e-mail message by message. Eventually, it was reduced to approximately 2 gigabytes of archival records.

Realizing that the volume of e-mail would continue to increase and seeking to find a tool that would help in the timely transfer of e-mail to the State Archives, staff applied for and received a grant from NHPRC. Through this grant, the Preservation of Electronic Mail Collaboration Initiative, the State Archives, in conjunction with the staff of Pennsylvania Historical and Museum Commission, Bureau of Archives and History and the KDLA, developed and tested an e-mail collection and preservation tool called EMCAP [17]. The concept for the tool was developed, in part, from requests for an easier way to transfer e-mail records to the Archives. The State Archives is testing the tool with the Office of Community and Citizen Affairs in the Governor’s office, the Corporations Division in the Secretary of State’s office, and director-level positions in the State Emergency Management Office within the Department of Crime Control and Public Safety. Each agency utilizes a different e-mail system — Microsoft Exchange, Novell GroupWise, and IBM Lotus Notes. In the EMCAP scenario, the user determines which e-mail messages are records of archival value per his or her agency’s records retention schedule, and, using their email client, deposits those e-mails by “dragging and dropping” them into an EMCAP folder. The EMCAP e-mail account and the folders (created by the user) within that account physically reside on a server located at the State Archives rather than on an agency machine or a machine at ITS. An original copy of each e-mail is saved and the e-mail header information and text is converted to XML. Depending on the size of the attachment, it is either kept with the message itself or is de-coupled and a relative link back to it is added to the message. To date, the pilot has ensured the retention of over 38,000 e-mail messages at the State Archives.

2.2 Providing Access (Short and Long-Term)

2.2.1 Websites
The State Library and the State Archives jointly developed and manage a web interface for accessing the web pages harvested with Archive-It. Because the service allows for only minimal metadata entry (and only at a high level), keyword searches are drawn from indexes based on the text found on websites or other objects. And while the State Library and State Archives do not currently have preservation tools to ensure long-term access to all of the harvested material, the files are stored in preservation-friendly ARC and WARC files and the Internet Archive manages and stores this content on their servers in California, Amsterdam, and Paris. The State Archives also annually purchases hard drives of the content and stores them locally.

2.2.2 Publications
The State Library has developed a separate web access tool — utilizing CONTENTdm — for its digital repository of publications. As part of a pilot training program offered to state agency staff in 2008 and increased outreach efforts by State Library staff, state agency personnel responsible for publishing or
managing agency-created publications have been informed about the new repository service and encouraged to submit their publications as a starting point to answer both their short-term access and long-term storage needs. State Library staff is hopeful that this educational effort will increase agency participation in the use of the repository, in terms of both submissions and access. Access to information within the repository is fairly straightforward as CONTENTdm allows for two levels of search: item-level metadata records and full-text. And, while “true” preservation management tools are not currently in place, the State Library is storing digital masters and access copies locally on a departmental storage area network (SAN), and duplicates are stored off-site at OCLC’s Digital Archive in Dublin, Ohio, where multiple copies of the files are created, managed, and stored [18].

Driven by the desire to offer a single point of access to all of the State Library’s digital content, the State Library subscribes to WebFeat’s federated search tool. Presently, WebFeat searches the Library’s catalog, digital repository, and all of the Library’s electronic journal subscriptions [19]. The State Library is also investigating the feasibility of adding harvested web content from Archive-It to the digital repository, thus making that data searchable through WebFeat, as well.

2.2.3 Geospatial Data
Access to North Carolina geospatial data is typically provided by either its county of origin or via CGIA’s NC OneMap. “Archival” programs are not as well established; the fact that GIS data exists in many different places — held by creators, various users who have downloaded or requested it, and as part of CGIA’s NC OneMap — means that its chances for long-term survival are increased. However, sustainable access is an issue that must be addressed. One goal of the State Archive’s GeoMAPP initiative is to build the infrastructure and knowledge base regarding how to provide long-term access to the geospatial content that the State Archives will collect. Using funding from NDIIPP, the State Archives will purchase storage to begin collecting geospatial data. A mirror site for CGIA is also being considered. And, instead of building another access tool, the GeoMAPP partners are hoping to leverage NC OneMap to access this data.

2.2.4 E-mail
The goal of the EMCAP grant was to develop and test a tool for the collection and preservation of e-mail. “Archival” programs are not as well established; the fact that GIS data exists in many different places — held by creators, various users who have downloaded or requested it, and as part of CGIA’s NC OneMap — means that its chances for long-term survival are increased. However, sustainable access is an issue that must be addressed. One goal of the State Archive’s GeoMAPP initiative is to build the infrastructure and knowledge base regarding how to provide long-term access to the geospatial content that the State Archives will collect. Using funding from NDIIPP, the State Archives will purchase storage to begin collecting geospatial data. A mirror site for CGIA is also being considered. And, instead of building another access tool, the GeoMAPP partners are hoping to leverage NC OneMap to access this data.

As in so many other environments in the public and private sectors, agency staff indicates that they feel overworked and do not have time to incorporate new and complex actions into their existing workflows. This is another challenge to implementation and sustainability of digital information management programs. The State Library and State Archives must ensure that the actions necessary to participate are not burdensome on any one individual, can easily be incorporated into existing workflows, and will produce a reportable benefit that management will appreciate and encourage. While the first two of these requirements are difficult to achieve, the third seems nearly impossible in the short term. However, as technology continues to advance and stored digital information becomes obsolete or disappears altogether, the value will become obvious. Some attempts to consolidate data management have been made by ITS, the central service point for network and application support for many (but not all) state agencies. But push-back from

8 Until such an access system is developed, patron requests are filled by burning a CD of the e-mail account in its original file format. It is incumbent upon the researcher to have the tools to read the messages.
agencies has been strong as ITS is seen as taking on too many initiatives without sufficient staffing, resources, or a fundamental understanding of all the issues required to succeed. However, the North Carolina state legislature has recently encouraged ITS’ involvement and allocated funding for new projects. In the summer of 2008, the legislature passed, through HB 2436, S.L. 2008-107, appropriations to develop a detailed plan to implement the recommendations contained in the Geographic Information System Study, including a cost study to centralize the management of all GIS resources, projects, and – ideally – long-term preservation of the products of this undertaking by NC ITS [section 6.13]. Similar plans are in the works for e-mail [section 6.14].

The challenges are great — while the investigation of long-term maintenance needs for all data continues, and forging inter-agency partnerships is daunting, some progress has been made. NC OneMap, the State Library’s Digital Publications Repository, the State Archives’ e-mail preservation tool, and the State Library and State Archives’ web harvesting efforts are all practical examples of how consolidation can work. Continued participation in grants that allow testing and research of newly developed tools that ease the burdens of digital curation across the spectrum of state government agencies are sure to build upon these successes in the near future.

4. ACKNOWLEDGMENTS

Our thanks go out to all of the organizations that we collaborate with, including the Internet Archive, North Carolina State Publications Clearinghouse, Washington State Archives, National Historical Publications and Records Commission, Library of Congress, Data Intensive Cyber Environments group, Renaissance Computing Center, Center for Geographic Information and Analysis, Kentucky Department for Libraries and Archives, Kentucky Commonwealth Office of Technology’s Division of Geographic Information, Utah State Archives, Utah Automated Geographic Reference Center, Smithsonian Institution Archives, Rockefeller Archives Center, Pennsylvania State Archives, North Carolina Governor’s Office, North Carolina Secretary of State’s Office, and North Carolina Department of Crime Control and Public Safety.

5. REFERENCES

[8] NC OneMap: Geographic Data Serving a Statewide Audience. DOI= http://www.nconemap.com
[17] E-mail Collection and Preservation tool. DOI= http://www.records.ncdcr.gov/EmailPreservation
[18] OCLC. Digital Archive. DOI= http://www.oclc.org/digitalarchive
Invited Paper: MoReq2: a European Contribution to the Preservation of Electronic Records

Marc Fresko
Inforesight Limited
Kenley, CR 5HD UK
marc.fresko@inforesight.co.uk

ABSTRACT
MoReq and its successor MoReq2 are European model specifications of requirements for Electronic Records Management systems, often referred to as ERM or EDRM systems. There are several other specifications that set out to define model requirements for EDRM systems – notably the US DoD 5015.2 standard. However, the other specifications almost all are designed expressly for government bodies in one country1; MoReq and MoReq2 are differentiated in three ways:

1. they are designed to apply to all sectors (public, private and not-for-profit alike);
2. they apply to all members states of the European Union;
3. they include features that have been found to be valuable in practice, even though they are not strictly required for the theoretical management of records.

This paper concentrates on the third differentiator, and specifically on features that address digital preservation, notably:

- migration;
- “components”;
- automated rendition;
- import and export;
- preservation metadata;
- XML schema.

General Terms
Management, documentation, design, reliability, standardization, languages.

Keywords
MoReq2, Model Requirements for the Management of Electronic Records, preservation, digital preservation, sustainability, records, electronic records, European Commission, XML.

1. BACKGROUND: MoReq AND MoReq2
1.1 Records
This paper makes frequent reference to “records”. The formal definition of “record” is given in the international standard that defines Records Management [7] as:

“Information created, received, and maintained as evidence and information by an organization or person, in pursuance of legal obligations or in the transaction of business”.

In less formal terms, “records” are documents that tell us what an organization (any organization, in government or in commerce) has done, how decisions were reached, what transactions have been executed, and so on. Importantly, records are not limited to those documents that are bundled up and sent to an archive; they include also those documents that are used every day to manage the organization.

1.2 Electronic Records Management
Records Management is a distinct discipline with a long lineage. Its history in Europe can be traced back to the Medieval era. And we have lived with the need to keep electronic records2 ever since computers were first used in commercial applications3, though in practice many organizations chose to keep paper or microform copy records instead.

The early history of keeping electronic records, and indeed of making them accessible over a long period, is trivial, as it applies to highly structured, or transactional data. Banks, insurers, manufacturers, airlines and others have routinely created, managed and (wherever necessary) preserved such structured electronic records without undue difficulty, without complaint, without massive loss (usually!), and certainly without multi-million dollar academic research programs. This is because these

---

1 The singular exception is the recently-published specification from the International Council on Archives [6]. It is too soon to judge acceptance and impact of this publication.

2 For reasons lost in history, we refer to “electronic records” instead of “digital records”, even though we refer to “digital preservation” and not “electronic preservation”. In principle we should prefer the term “digital records” to distinguish computerized records from “analogue electronic records” such as old audio recordings; in practice, we don’t.

3 The first computer known to be used in a commercial application was “LEO”, Lyons Electronic Office I, in the United Kingdom in 1951. This 500kHz machine, armed with 2048 words of ultrasonic mercury delay line memory, was used for several applications including valuations, payroll and inventory management.
early electronic records were in formats that were straightforward and understood by (often designed by) the institutions in question. However, electronic records rapidly became at once more complex and more widespread. The complexity arose because of the evolution of ever-more capable software for word processing, spreadsheets, project management, graphics, presentations and so on; and to compound matters, the formats used mostly were not only proprietary but secret. The widespread usage resulted from a series of technological waves – first office computing in general and PCs in particular, then e-mail, and then the web. We now all know that virtually all records of enduring value are created, edited, used, modified, and stored electronically. Some records are still created and stored in paper form – correspondence from citizens and retail customers being the prime example – but they now constitute only a tiny fraction of the records created today. Combine with this the very real practical difficulties of storing paper copies of electronic records, the need to manage electronic records is immediately apparent – and thus the need to preserve them also.

Electronic records do not manage themselves, and certainly do not preserve themselves. Without active management, all manner of problems arise, from the change and deletion of records that ought to be fixed, through the inability to find records, to loss of access because of technological evolution.

All this gives rise to a need for systems that can manage electronic records so that they can be used on a daily basis as the information lifeblood of the organization – Electronic Records Management (ERM) systems. Because it turns out in practice to be convenient for the same systems to manage documents that are not formal records (mainly because many of them are destined to become formal records) these systems are also referred to as Electronic Document and Records Management (EDRM) systems.

The need for a formal standard expression of the capabilities of ERM dates back to the “R/DIM” initiative in Canada in 1996, a work which itself grew from a Canadian government initiative of the early 1990s. The resultant RDIM specification was widely read and must have been influential, but is rarely cited and difficult to find today.

More significant by far is the US Department of Defense’s standardisation, usually referred to by its number as “5015.2” [11]. Catalyzed by problems observed with the management of electronic records of the first Gulf war, it was first published in 1997, and updated several times since then (most recently in 2007). It has been extremely influential on both the capabilities of the software industry and in fostering the expectation of the user base. The former has engineered literally dozens of software application to comply with the various versions of the standard, and the latter – at least in North America – demand compliance as a demonstration of the ability to manage electronic records.

Similar pressures in other countries gave rise to comparable specifications in other countries, including the “PRO” specification [10] in the UK (by far the most influential of this set), DOMA in Germany, and many others in countries as far apart as New Zealand and South Africa. All of these specifications, without exception, were developed by organizations of national government to serve as guidance for other government bodies.

1.3 The MoReq Specification

The MoReq specification [2] is similar in concept to the other specifications. However, its genesis was different. It was conceived at an international level, to be applicable across national boundaries throughout the European Union; and it was from the outset intended to apply to all sectors, government and otherwise.

MoReq was conceived by the DLM Forum, an international organization concerned with archives and records management. Its initials “DLM” originally stood for the French words “Données Lisibles par Machine” – “Machine Readable Data” in English. However, in a bizarre reversal of acronym causality, the Forum voted in 2003 to change the meaning of DLM to the more anglophone “Document Lifecycle Management”.

The Forum convinced the European Commission to fund MoReq’s development, and as a result MoReq was produced by a team of consultants led by the author of this paper. It was published in 2001, and immediately became a success across Europe. Its success can be measured by the number of translated versions produced around the world: we have identified 8 full translations and a handful of adaptations, some from countries well outside of Europe (South America, Asia). It was also used in the USA, for example at Indiana University.

MoReq differed from the other specifications. Because it was written by consultants, it took into account a wide range of experiences with both EDRM systems and with procurement specifications. Accordingly, observers viewed it as easier to read and understand than the others. More importantly, it also includes many practical requirements – features that are not strictly essential for a records management in theory, but which were deemed by the authors to be important in practical office settings. A such set of features was an early set of requirements for digital preservation.

1.4 The MoReq2 Specification

For all its rapid international success, MoReq was far from perfect. There were weaknesses in its metadata model; it was not maintained and hence lost its currency; and it lacked any external governance or compliance testing structures. At the same time, and driven partly by the enlargement of Europe, there was growing demand for a “better and bigger” version of MoReq, one

\[4\] The brief summary that follows is restricted to materials published in English. The author is not aware of earlier works in any other language.

\[5\] The R/DIM specification appears not to have survived on the internet except in the Internet Archive [1]. Interestingly, despite its age, it includes (in its requirements 3.3.2 to 3.3.8) features that relate specifically to digital preservation that are absent from most of the more recent specifications.

\[6\] See http://dlmforum.eu.

\[7\] Brazilian Portuguese (two versions), Dutch, Polish.
with an ongoing management regime. Accordingly, the DLM Forum turned anew to the European Commission, which again agreed to fund development.

So it was that MoReq2 was developed, during 2007, again by a team of consultants. The author of this paper again was in the lead, this time opting for a highly consultative process involving over 200 experts, interested parties and organizations from around the world. The finished MoReq2 was published in early 2008. Like its predecessor, it is intended to cater for all kinds of organization, in any European country. Longer and more detailed than any other specification, it also contains many practically-important requirements that are absent from other specifications, such as automatic numbering, exception processing and the like.

Importantly, MoReq2 was prepared with the support of most of the ERM/EDRM industry. Just some of the 44 companies who explicitly expressed their support are listed in Table 1.

Table 1: Selected vendors supporting the MoReq2 project

<table>
<thead>
<tr>
<th>Adobe</th>
<th>Meridio (now part of Autonomy)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CA</td>
<td>Microsoft</td>
</tr>
<tr>
<td>Capgemini</td>
<td>Objective Corporation</td>
</tr>
<tr>
<td>EDM Solutions</td>
<td>Open Text Corporation</td>
</tr>
<tr>
<td>EMC</td>
<td>Oracle</td>
</tr>
<tr>
<td>FileNet (now part of IBM)</td>
<td>SAP</td>
</tr>
<tr>
<td>Fujitsu</td>
<td>Tower Software UK (now part of HP)</td>
</tr>
<tr>
<td>Getronics</td>
<td>IBM</td>
</tr>
<tr>
<td>IBM</td>
<td>Xerox</td>
</tr>
</tbody>
</table>

Like its predecessor, MoReq2 has prompted many translation projects: the French and Russian translations have already been published, with translations into Catalan, Korean, Romanian, Slovenian, Spanish and other languages well advanced.

It consists of the following publications:

- the specification itself [3];
- a comprehensive metadata model [3];
- a testing framework (test conditions, expected results) [4];
- an XML schema [5].

2. DIGITAL PRESERVATION FEATURES IN MoReq2

The general philosophy of MoReq2 is to include features that are necessary in a practical EDRM system, even if they are not strictly necessary for the management of electronic records. One set of such features addresses digital preservation. The features are described briefly here.

2.1 Migration

First, MoReq2 includes requirements for the migration of specified records to new formats, to support active management efforts for digital preservation. These requirements are mandatory.

2.2 Recognition of “Components”

Possibly the most significant novel feature of MoReq2 is its recognition of electronic objects that are smaller than electronic records. For example, web pages tend to be built up of text files, image files, style sheets etc., each of which is treated as a separate entity by any operating system, but all of which have to remain tightly linked together in an EDRM system if the integrity and usability of the records are to be maintained. MoReq2 coins the term “component” to describe these objects, in the absence of any agreed term.

MoReq2 contains a sophisticated conceptual model to relate the various electronic and physical entities that make up a set of records stored in an ERM system. The model, expressed as an entity-relationship diagram, is a development of a similar model in MoReq, but is much more rigorous and powerful. A small extract from this model, in Figure 1, shows how components are related to records.

![Figure 1: Relationship of components and records](image)

This shows that all documents and records are made up of at least one component; and that some are made up of several. So, for example, a simple spreadsheet is made up of only one component; but a record consisting of linked spreadsheets is made up of several components. This paper could consist of a single object (if the graphics are contained entirely inline) or could be several components if the graphics are implemented by means of embedded links to external images.

Components play a critical role in digital preservation. To continue the example of websites, if it is necessary to preserve web pages made of many components (as almost all pages are) then it is the components themselves, and their links, that will need to be migrated – not the page “object” itself.

MoReq2 recognizes the need to manage components in a way that preserves integrity and usability. Uniquely, it also specifies requirements for migration. For example, if a website is made up of an HTML component plus GIF and PNG components, it might become necessary to migrate the GIF components (and nothing else), say from GIF format to PNG format. However, doing this would require the HTML tags that reference the GIF files to be changed to reference the new PNG files. This is unavoidable, but very counter-intuitive to Records Managers who (rightly) believe that the content of records must not be changed. The scenario of component migration is a rare situation in which a change to content (changing filenames from GIF to PNG in this example) is

---

9 For example, the home page of the NARA website in early March 2009 is made up of one HTML file, 3 GIF files, 2 JPEG files, 7 Javascript files and 12 cascading style sheets.
a prerequisite for the maintenance of usability. MoReq2 contains several requirements that explain how migration processes need to take this into account.

2.3 Automated Rendition
MoReq2 requires that systems have the ability to render all electronic records to a preservation format of choice at the time of capture. So, for example, if PDF/A [8] is selected as the preferred preservation format for “four-cornered” records, then a compliant system would have to be able to render, automatically, all appropriate documents (so not audio, video etc.) into PDF/A format as soon as they are captured, subsequently managing both the original and PDF records in parallel. This relatively simple feature, which we believe to be totally novel to standard specifications, will provide a major boost to efforts to build repositories of preservable and accessible electronic records.

2.4 Import and Export Features
MoReq2 contains detailed requirements supporting the import and export of electronic records. Unlike most other specifications, the requirements explain in detail how not just the records themselves are involved, but also their metadata and audit trails. Details covered include error handling (in the event of duplicates or incomplete data structures – both major concerns given the complex hierarchical structures associated with records management) and protocols to confirm successful exchange. These features can be useful to export records from an ERM or EDRM system to either a preservation system or a newer system.

2.5 Preservation Metadata
MoReq2 contains a uniquely detailed metadata model [3]. It consists of 158 different metadata elements, each cross-referenced with the appropriate entities to which it can apply (component, record, file, class etc). In effect this defines 345 distinct metadata elements – more than any other electronic records management specification. The elements are chosen to provide all the metadata needed for full Mooreq2 compliance.

Not only is the metadata model extensive, it is also detailed. Each element is described in some depth, in the form of a table. An example is shown in Figure 2 (note that the name of the element is structured to comply with ISO 23081 [9]).

The example in Figure 2 is one of the simpler elements. Other metadata element descriptions specify not just the definition of the element, but how it is used, how it relates to other elements, and in many cases from where its values may come, how they are derived, and what values are valid. Contrast this with a comparable definition from DOD 5015.2, shown in Figure 3:

![Table of a MoReq2 metadata element](image)

Figure 2: Example of a MoReq2 metadata element

2.6 XML Schema
The final piece of the jigsaw is the MoReq2 XML schema [5]. This expresses the metadata model in XML, and provides – hopefully – an unambiguous mechanism to allow the exchange of electronic records, including their metadata and audit trails, between systems. At this stage, it is just “hopefully” as there is no credible mechanism to test the correctness and completeness of the schema itself. However, we are hopeful that, as soon as software claiming to comply with MoReq2 becomes available, it will be possible to test the schema at least partially.

3. THE FUTURE OF MOREQ2
We already know that all of the major ERM/EDRM system vendors have MoReq2 compliance projects under way. Already, buyers in Europe are specifying MoReq2 compliance as (in some cases) mandatory. The near future will most likely see MoReq2-compliant products reach the market, first in Europe then beyond.

The original MoReq specification was very popular, but unfortunately it suffered from a lack of governance. Once published, MoReq had no ongoing supervision or management at all – it effectively was orphaned.

This time around, the DLM Forum has set up a solid governance structure to manage MoReq2 on an ongoing basis. It proved challenging to get this started, due in part to the fact that the governance is to be provided on a voluntary basis; but early indications are hopeful. The new MoReq2 Governance Board (MGB) will:

- oversee the software compliance testing regime;
- ensure the “brand” names MoReq and MoReq2 are not misused, and in particular ensure no false claims of compliance are published;
- monitor the accuracy and acceptability of translations and of the localization chapters;
- look after the maintenance and ongoing development of MoReq.

The MGB is made up of senior representatives of institutions from several countries with a strong interest in MoReq2. They are listed in Table 2.
4. LIMITATIONS?
Following the MoReq2 specification will not produce an archival management system. Nor will it produce a digital preservation system. What it will produce is a system that is useful to manage electronic records on an everyday basis – and one which also provides basic facilities to support active digital preservation management.

5. WHAT NEXT?
MoReq2 is not an end point in a trajectory, and is definitely not an end in itself. Rather, it is a point in a journey. As we all learn more, so it improvements will be made to MoReq. Probably the most important improvement we can think of is the integration of MoReq2 with other initiatives – DoD 5015.2 in the USA, perhaps VERS in Australia, certainly ISO 15489. There really is no plausible reason for this to continue as a solely European venture – and there is even less justification for the existence of different electronic records management standards on different continents. To be sure, uniting these different ideas would be challenging – but a challenge well worth winning.

6. REFERENCES
Digital Curation of Humanistic, Multimedia Materials: Lessons Learned and Future Directions

Megan Winget, Caroline Frick
University of Texas
School of Information
Austin, TX 78712 USA
{megan,cfrick}@ischool.utexas.edu

Jerome McDonough, Allen Renear
University of Illinois
Graduate School of Library and Information Science
Champaign, IL 61820 USA
{jmcdonou,renear}@uiuc.edu

Henry Lowood
Stanford University
Stanford University Libraries
Stanford, CA 94305 USA
lowood@stanford.edu

ABSTRACT
Library and archival literature often points to the audio-visual and other non-text preservation communities as leaders in envisioning new, digitally driven curation methods and practices. A closer look, however, reveals that the majority of such institutions and key practitioners — both commercial and not-for-profit — remain committed to 19th century conservation theory and models. This panel features interactive case studies that illustrate participants’ practical and theoretical experiences with multimedia and new media collections, and will provide alternative approaches to traditional models of appraisal, collection development, access, and preservation. In this panel discussion we will discuss what it means to curate a collection of multimedia and interactive media, how such collections might be used in the humanities, and what role curators play in creating, preserving, and promoting their use.

Panel Participants:
Caroline Frick, Assistant Professor, University of Texas at Austin, School of Information & Department of Radio, Television & Film. Director, Texas Archive of the Moving Image
Henry Lowood, Curator for History of Science & Technology Collections and Film & Media Collections, Stanford University Libraries. PI, How They Got Game Project
Jerome McDonough, Assistant Professor, University of Illinois at Urbana-Champaign, Graduate School of Library and Information Science.Co-PI, Preserving Virtual Worlds Project
Megan Winget, Assistant Professor, University of Texas at Austin, School of Information. PI, Creation Processes and Artifacts of the Game Industry
Allen Renear (Moderator), Associate Professor, University of Illinois at Urbana-Champaign, Graduate School of Library and Information Science. PI, Extending Data Curation to the Humanities: Curriculum Development & Recruiting

This work is licensed under the Creative Commons Attribution-Noncommercial-No Derivative Works 3.0 Unported license. You are free to share this work (copy, distribute and transmit) under the following conditions: attribution, non-commercial, and no derivative works. To view a copy of this license, visit http://creativecommons.org/licenses/by-nc-nd/3.0/.

DigCCurr2009, April 1-3, 2009, Chapel Hill, NC, USA

Keywords
Digital curation, non-text, multimedia.

1. INTRODUCTION
Digital curation means many things to many people. In the scientific community, the term refers to the act of collecting, describing, and providing continued access to data generated from research [2], [7]. In the humanities and social sciences, digital curation tends to be viewed as an extension of archival practice, encompassing appraisal, organization, description, and preservation activities [1]. Much of the current research focuses on scientific data (c.f., [3]; [4]), and the development of institutional repositories (c.f., [6]; [5]). Very few researchers have provided insight into digital curation practice in terms of humanistic, specifically non-text and multimedia materials. This oversight is unfortunate, as these materials present general challenges to traditional Information Science theory and practice, and provide the opportunity to augment digital curation theory.

Some specific challenges central to the practice of digital curation, and which multimedia materials present:
• Sustained Involvement with Creators
• Representation of Complex Objects
• Providing Sustained Access
• Collaborative Curation and Access

Panel participants will relate their experiences in creating, managing, and providing sustained access to large collections of non-text and multimedia materials, and will address the changing roles and expectations of information professionals as pertains to digital curation.

2. DISCUSSION TOPICS
2.1 Sustained Involvement with Creators:
Megan Winget: Creation Processes and Artifacts of the Video Game Industry

The primary goal of this research project is to come to a better understanding of the video game industry’s creation methods, behaviors, and attitudes for the purpose of building more meaningful models of preservation and collection of these materials. Most of the current preservation projects for new media and video games focus on the end products: the “final” art objects
or the released video games themselves. Within the archival community, however, there is the realization that digital preservation starts with creation; it is impossible to reliably and authentically preserve an object without having a very good idea of the circumstances and particulars of that object’s creation. This project team hopes to shed light on those conditions. We believe that the products of this research project will support better collection, access and preservation of these significant and important cultural artifacts.

2.2 Representation of Complex Objects: Jerome McDonough: Preserving Virtual Worlds
Archiving games requires the storage and organization of vast amounts of data and metadata. While there have been efforts, such as the OAIS Reference Model and the affiliated XML Formatted Data Units, to developed standardized models and mechanisms for archiving digital information, these have focused on preserving digital data, not software. The Preserving Virtual Worlds project is a multi-institution collaboration between the University of Illinois, Rochester Institute of Technology, Stanford University and the University of Maryland investigating the preservation of computer games and interactive fiction. We will report on our efforts to develop mechanisms to track the wealth of data and metadata needed to preserve computer games, and the problems we have encountered in trying to develop those mechanisms in a way which aligns with existing and developing standards with the library and data curation communities.

2.3 Collaborative Curation and Access: Caroline Frick: Texas Archive of the Moving Image
Historical societies, libraries, and archives frequently struggle with providing access to “alternative” formats such as motion picture film and video. Such audiovisual materials possess their own set of preservation problems, and the obsolescence of playback equipment causes large numbers of archival content to remain hidden from view. Working alone, organizations lack the budget as well as technical expertise to grapple with these materials. Furthermore, traditional museological approaches to prioritizing the preservation of the archival object present serious challenges to statewide digitization initiatives. The Texas Archive of the Moving Image's (TAMI's) “Film Round-Up” program is a collaborative attempt to discover, digitize, and curate such materials with members of the public as well as organizations across the state. Audio-visual content digitized as a result of this initiative is presented online via TAMI's library: http://www.texasarchive.org/library. A contemporary case-study, the Round-Up program illustrates the challenges of collaborative curatorial practice learned "on the road."

2.4 Collection Development and Interpretative Activities: Henry Lowood: How They Got Game Project
Henry Lowood will speak from experience gathered from projects undertaken by the How They Got Game project, which started in 2000. The focus in collection development will be the collection and exhibition of collections in the history of interactive software, especially digital games, simulations and virtual worlds. An important aspect of this topic will be the creation of collections of historical documentation that supplement the preservation of interactive media in its various forms by capturing the history of events and activities in games and virtual worlds. Examples of the latter activity include the Archiving Virtual Worlds collection and the Machinima Archives, both undertaken in collaboration with the Internet Archive. As an example of interpretive activities derived from the collections, Lowood will describe “Game Capture,” a work created for the Australian Centre for the Moving Image.

3. PANEL STRUCTURE
This panel will be structured to maximize conversation among the panelists and with the audience. Each panelist will be given time to present their work and experience curating collections of multimedia and interactive media, how such collections might be used in the humanities, and what role curators can play in the creation, preservation, and promotion of their use. A moderated discussion and opportunity for audience Q&A will follow panelists’ presentations.

4. REFERENCES
Invited Demo Session I: Digital Curation Tools and Demos

Andreas Rauber
Vienna University of Technology
A-1040 Vienna, Austria
rauber@ifs.tuwien.ac.at

Jeffrey van der Hoeven
Koninklijke Bibliotheek
2595 BE, The Hague,
The Netherlands
Jeffrey.vanderhoeven@KB.nl

Raymond J. van Diessen
IBM Business Consulting Services
1066 VH, Amsterdam,
The Netherlands
Raymond_vandiessen@nl.ibm.com

Richard Pearce-Moses
Arizona State Library, Archives, and Public Records
Phoenix, AZ 85007 USA
rpm@lib.az.us

Heather Bowden
School of Info. & Library Science
Univ. of North Carolina at Chapel Hill
Chapel Hill, NC 27599 USA
heather@longtermdata.com

Jeffrey Pomerantz
School of Info. & Library Science
Univ. of North Carolina at Chapel Hill
Chapel Hill, NC 27599 USA
pomerantz@unc.edu

ABSTRACT
This invited demonstration panel session brings together an international collection of tools for performing and facilitating digital curation in the practice setting, as well as for use in the education setting. This is the first of a two-part tools and demo session. Tools to be demonstrated in this session include: Plato, the Planets preservation planning tool [1]; Hoppla (Home and Office Painless Persistent Long-Term Archiving) system [2]; Dioscuri, a modular emulator [3]; the Universal Virtual Computer (UVC) [4]; the Preservation Manager [5]; PeDALS (Persistent Digital Archives and Library System) [6]; and the DCE (Digital Curation Exchange). Additionally, two invited papers included in these Proceedings provide further information on Plato and Hoppla.

Keywords
Digital libraries, digital curation, data curation, digital preservation, planning.

1. REFERENCES
[4] Digital Asset Preservation Tool,
http://www.alphaworks.ibm.com/tech/uvc
[5] Preservation Manager,

This work is licensed under the Creative Commons Attribution-Noncommercial-No Derivative Works 3.0 Unported license. You are free to share this work (copy, distribute and transmit) under the following conditions: attribution, non-commercial, and no derivative works. To view a copy of this license, visit http://creativecommons.org/licenses/by-nc-nd/3.0/.

DigCCurr2009, April 1-3, 2009, Chapel Hill, NC, USA
Invited Demo: Creating a Preservation Plan Using the Preservation Planning Tool Plato

Hannes Kulovits, Christoph Becker, Michael Kraxner, Andreas Rauber
Vienna University of Technology, Karlsplatz 13, 1040 Vienna, Austria, +43/(0)1/58801-0
kulovits, becker, kraxner, rauber @ ifs.tuwien.ac.at

ABSTRACT
The rapid technological changes in today’s information landscape have turned the preservation of digital information into a pressing challenge. A lot of different strategies, i.e. preservation actions, have been proposed to tackle this challenge. However, which strategy to choose, and subsequently which tools to select to implement it, is a non-trivial task. Creating a concrete plan for preserving an institution’s collection of digital objects requires the evaluation of available tools against clearly defined and measurable criteria. Preservation planning aids in this decision making process to find the best preservation strategy considering the institution’s requirements, the planning context and possible actions applicable to the objects contained in the repository. Performed manually, this evaluation of possible solutions against requirements takes a good deal of time and effort. In this demonstration, we present Plato, an interactive software tool aimed at supporting institutions in the process of creating preservation plans.

General Terms
Design, Experimentation, Measurement, Standardization

1. INTRODUCTION
The fast changes of technologies, file formats and information systems have considerably shortened the lifespan of digital objects. Digital preservation has become a pressing challenge which has been further underlined by a recent survey among professional archivists [6]. A lot of different strategies, i.e. preservation actions, have been proposed to tackle this challenge: migration and emulation are the most prominent ones. However, even the decision to either apply migration or emulation is quite hard, and it gets more complex when it comes to pick a migration tool or specific emulator. The process of evaluating potential solutions against specific requirements and building a plan for preserving a given set of objects is called preservation planning. Presently this is a mainly manual process with little tool support. The main goal of the software tool Plato is to support the planner in the complex process of creating a preservation plan for a certain collection to be sustained into the future.

2. PLATO
The preservation planning tool Plato\(^1\) [2] implements the PLANETS\(^2\) Preservation Planning approach [5], which provides a solid way of making informed and accountable decisions on which solution to adopt in order to optimally preserve digital objects for a given purpose. It defines measurable requirements for preservation strategies in a hierarchical form and evaluates them in a standardised setting to arrive at a recommendation for a specific solution. This is done by applying preservation actions to a representative set of digital objects in an experimental setting and evaluating the outcomes. The procedure can be applied for any class of strategy, be it migration, emulation or any other approach. The PLANETS preservation planning workflow consists of four phases, each concerned with particular aspects of the planning activity: (1) Requirements definition, (2) Evaluation of potential strategies, (3) Analysis of results, and (4) Building a preservation plan.\(^2\) The end result of the workflow is a preservation plan which contains a description of the context and the decision taken, including the complete

\(^1\)http://www.ifs.tuwien.ac.at/dp/plato
\(^2\)http://www.planets-project.eu
A core element of Plato is the explicit definition of preservation requirements. Figure 2 shows the requirements for a preservation endeavour of a web archive laid out in a tree structure. The definition of requirements in a tree structure is often done in a workshop setting using mind-mapping software\(^6\). The mind-map created during the workshop can directly be imported into Plato. As each requirement must be measurable the tool offers a fully flexible way to specify a wide range of measurement scales.

The tool is integrated into the PLANETS Interoperability Framework based on open J2EE and web technologies. Through this environment it integrates registries and services for preservation action and characterisation through flexible discovery and invocation. Characterisation services such as DROID\(^4\) and JHove\(^5\) are used for format identification and property extraction; based on this information, applicable action services such as emulation tools or the migration services provided by CRiB [4] are discovered through available registries. Figure 1 shows chained migration services the CRiB registry can offer for migrating JPG images to different file formats. Each migration path can be selected to be considered as a potential preservation action and evaluated after it has been applied to the chosen sample objects. Alternatives for all preservation actions not available through registries may also be included in the evaluation.

The evaluation of experiments is probably the most complex and time consuming task of the whole workflow. Based on the outcome of the experiments, for each requirement the planner has to assess to which degree it is fulfilled. For example, if the line breaks in a document or the colour model of an image has been preserved accurately. The integration of the extensible characterisation languages (XCL) [3] enables automatic object comparison; all properties supported by XCL for a certain object type can be compared automatically, for instance image width or colour depth. Comparison and validation of objects as an essential feature of Plato maps the specified requirements such as essential object characteristics to measurable criteria that can be compared automatically. It thus considerably improves the repeatability, documentation and automation of preservation planning.

The applicability and usefulness of the tool has been validated in a series of workshops and case studies which involved various institutions [5, 1].

3. CONCLUSION

Until now, the process of preservation planning has predominantly been a manual and ad-hoc activity, demanding a lot of experience and effort. Plato is a software tool that implements a well-documented and validated preservation planning methodology and integrates services and registries for preservation action and characterisation. Features such as discovery of potential preservation action services through registries and automatic object comparison abstract much of the complexity out of the planning process and enable preservation planning also for less experienced users. Current and future work is aimed at further integration of service registries to offer the planner a wider range of preservation actions, such as GRATE\(^6\) for emulation.

Acknowledgements

Part of this work was supported by the European Union in FP6, IST, through the PLANETS project, contract 033789.

4. REFERENCES


\(^{4}\)http://hul.harvard.edu/jhove
\(^{5}\)http://planets.ruf.uni-freiburg.de/
\(^{6}\)http://freemind.sourceforge.net
\(^{6}\)http://droid.sourceforge.net
ABSTRACT
Small businesses (small office/home office, SOHO) have tremendous amounts of digital information. At the same time, they have little to no expertise on how to manage it, not to mention caring for their long-term preservation, as even simple back-up strategies pose already drastic challenges.

This demo presents the Hoppla archiving system\textsuperscript{1} to provide digital preservation solutions specifically for small institutions and offices. It hides the technical complexity of digital preservation challenges by providing automated services based on established best practice examples. Appropriate preservation strategies and required tools for the collection are delivered via a web service, effectively outsourcing the required digital preservation expertise.

General Terms
Digital Preservation, Long Term Access, Archiving, SOHO

1. HOPPLA CONCEPTS
Hoppla\textsuperscript{2} - (Home and Office Painless Persistent Long-term Archiving) presents a new approach to automated digital preservation systems that are suited to their needs. Requirements for digital preservation of holdings in small office settings differ from those in professional settings caused by different levels of expertise and skills of the users, the different environments, and objectives. The requirements in automation of the archiving process are higher for an archiving software solution for users with less expertise in preserving and managing collections. In institutional settings, critical decisions in preservation endeavours can be made by skilled staff. Existing open source digital repositories, such as Fedora\textsuperscript{3} and DSpace\textsuperscript{4}, are useful environments for professional archiving, but usability and required knowledge for configuration and use do not meet the skills of user in small institutions and SOHOs. The requirements in data management as well as managing the preservation task fully automatically pose significant research challenges. Solving these will represent an important step forward in enabling non-experts to preserve their digital data in very much the same way and at acceptable levels of quality as they can currently preserve conventional objects for their own needs.

The underlying principle of the system is finding a best effort solution with respect to the available technology and skills of the users. We cannot assume a highly sophisticated computer environment; neither can we expect a profound knowledge in digital preservation or archiving. With Hoppla we are currently developing a solution that combines back-up and fully automated migration services for data collections in small institutions and SOHO settings. It will combine bit-stream preservation via LOCKSS-style back-up \textsuperscript{5} with logical preservation by automatically obtaining migration rules and tools. The system should provide the best available and most practical preservation solution.

The system builds on a service model similar to current firewall and antivirus software packages, providing a user-friendly handling of services, an automated update service and hides the technical complexity of the software. Data is ingested from a number of sources such as data carriers, email repositories and on-line storage locations, while back-up storage supports off-line and on-line storage media in both write-once as well as rewritable forms.

A detailed description of the concept of the Hoppla system is presented in [1]. This demo provides an insight into the current development of the system and challenges that an automated arching system is dealing with.

Figure 1 shows the basic architecture of the Hoppla system that is influenced by the OAIS reference model (ISO 14721:2003). It consists of four core components: acquisition, ingest, data management, preservation management and storage management. Two registries contain preservation rules and services. Both registries are updated automatically by an external update web service. The service

\begin{itemize}
\item \textsuperscript{1}Part of this work was supported by the European Union in the 6th Framework Program, IST, through the PLANETS project, contract 033789.
\item \textsuperscript{2}\url{http://www ifs tuwien ac at dp hoppla}
\item \textsuperscript{3}\url{http://www fedora info}
\item \textsuperscript{4}\url{http://www dspace org}
\item \textsuperscript{5}\url{http://www lockss org}
\end{itemize}
registrar} contains services and tools for object identification, characterisation, preservation, and preservation validation. The registrar also contains representation information about formats, for example the format specification. The preservation rule registry specifies preservation strategies for different types of objects.

The archiving system supports the acquisition of data from different sources via an API for acquisition plugins. The use of plugins allows to support all kinds of storage media and current as well as future data sources. The first version of Hoppla supports the disc acquisition and e-mail accounts. The selection of the digital objects to be preserved is performed in the ingest component. The module uses DROID\textsuperscript{6} to determine the object’s format. The ingest component creates a collection profile that is used to identify appropriate preservation rules for the collection.

The data management enriches the objects in the archive with metadata to ease later reuse. Metadata are created from the additional information captured by the acquisition component, the documentation of migration processes, and metadata extracted from objects. The Hoppla system is designed to deal with complex objects, objects that are represented by several files that are related to each other (e.g., web pages).

Preservation management controls the logical preservation of the objects. This means that it is responsible for performing migration strategies on the objects in its archive. To do this preservation tools and rules are requested from an update service. The web update service consists of two web services, a resource based registry and a dialog based service. The resource registry provides executable tools and services for migration and characterisation and metadata for object formats. The dialog based service provides preservation rules for the collection on the basis of functional requirements and constrains. The functional requirements include the collection profile and specific user requirements. The constrains for the preservation rules are primarily available storage and when indicated costs of preservation services or additional storage.

The application flow of Hoppla starts with the ingest of new objects from source media, the object’s formats are identified and a collection profile is created. Based on the collection profile suitable preservation rules are recommended by the web update service. For privacy reasons, the user can define the level of detail of the collection profile that is provided to the web update service. This way the user has strict control, which and how much metadata is sent to the server. New preservation rules and services are downloaded from the web service and preservation actions are performed on the client side. The new objects including the resulting objects from preservation actions are ingested into the collection and stored on storage systems.

The access module provides services that allow users to access the data stored in the archive. The access module further displays information about object dependencies and versioning history.

\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{hoppla_system.png}
\caption{Architecture of the Hoppla System}
\end{figure}

Storage management is responsible for bitstream preservation. The data provided by the data management component are stored on various storage media. The storage management supports multiple copies of the data, following the concept of the LOCKSS project. In order to store the data on various storage systems or media, storage management implements a reduced version of a storage resource broker\textsuperscript{7}. It provides a storage interface to access different storage systems, such as file systems or online storage system as well as write-once and rewritable media, by using plugins. All information and documentation generated by the Hoppla system are stored in XML format on the storage media.

2. RESEARCH ASPECTS

The first version of Hoppla focuses on basic migration, the web service for preservation rules and services, and the data management. A first version of the update service has been deployed providing preservation rules and services and an interface for administration. The functionality of the web update service will be further expanded and we specifically perform research on supporting the selection of preservation strategies considering functional requirements and constrains. Further improvement of the preservation rule selection based on the history of the collection will be done.

Another research aspect of the Hoppla system is context analysis of objects. The context of objects is essential for the interpretation of information objects and can provide enhanced access functionality to a collection. Various aspects of context in different dimensions can be automatically detected, and different views at multiple levels of granularity allow the extraction of the most appropriate connections to other digital objects. Information retrieval, data mining and OLAP-based approaches are used for structuring object collection and identifying relationships in multiple dimensions such a time, people, acronyms, content, etc.

3. REFERENCES


\begin{footnotes}
\end{footnotes}
Effective Access to Digital Assets: An XML-based EAD Search System

Junte Zhang  
Archives and Information Studies,  
Faculty of Humanities,  
University of Amsterdam  
j.zhang@uva.nl

Khairun Nisa Fachry  
Archives and Information Studies,  
Faculty of Humanities,  
University of Amsterdam  
k.n.fachry@uva.nl

Jaap Kamps  
Archives and Information Studies,  
Faculty of Humanities & ISLA,  
Faculty of Science, University of Amsterdam  
kamps@uva.nl

ABSTRACT
This paper focuses on the question of effective access methods, by developing novel search tools that will be crucial on the massive scale of digital asset repositories. We illustrate concretely why XML matters in digital curation by describing an implementation of a baseline digital asset search system that is fully XML-driven. The system aims to provide better access to archival material through digital finding aids in the Encoded Archival Description (EAD) standard. Relevant (parts of) archival descriptions within often lengthy and complexly organized digital archival finding aids can be found faster and with more ease. A succinct walk-through of the process of design and implementation of such a system is given, from a higher-level conceptual and generic view, where we start from the actual digital archival finding aid to the eventual delivery of the fonds to the user. Beyond this baseline, we propose a method for automatically providing extra archival context through automatic link detection between archival finding aids. We relate our efforts with the Encoded Archival Context (EAC) initiative.

Categories and Subject Descriptors
H.3.1 [Information Storage and Retrieval]: Content Analysis and Indexing; H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval; H.3.4 [Information Storage and Retrieval]: Systems and Software; H.3.7 [Information Storage and Retrieval]: Digital Libraries.

Keywords
Encoded Archival Description (EAD), archival access, information retrieval, information context, Encoded Archival Context (EAC).

1. INTRODUCTION
Digital curation is a recent umbrella term for a comprehensive approach to digital asset management [31]. The essence of digital curation is that it covers the whole live-cycle of a digital asset, from its creation to its future use. The comprehensive approach requires, on the one hand, activities centered on the digital assets (such as appraisal and selection, preservation, and records management), and on the other hand, activities centered on the future use (such as continual enrichment or updating, and effective access methods). The integration of both these aspects is a distinct characteristic of digital curation activities. In this paper, we will focus on the question of effective access methods, by developing novel search tools that will be crucial on the massive scale of digital asset repositories. These new search tools that are tailored to the data at hand, in our case a large collection of digital finding aids, are build from generic components. These search tools are not only valuable for online users but also for digital curators themselves, allowing them to better explore their repository and understand potential use of their digital assets. We illustrate concretely why XML matters in digital curation as our approach is fully XML-driven.

Archives, libraries and museums are memory institutions [9], which store the memories of societies, increasingly also digital assets, and enable their access. The archives have an important usage for users such as historians, as the archives offer primary sources (personal letter, handwritten diary, etc), which are used to reconstruct history. Historians are also the most respected users of archives [28]. These are described in archival descriptions, traditionally in paper form, so the creator or someone else can easier find them again. The archival material consists of records. A comprehensive overview of electronic record management is presented in [6] with the different ontological, epistemological and axiological points of view. An archival finding aid not only represents these records, but also their logical relationships and recorded information about the records, and this all makes an archive accessible.

The archival descriptions are increasingly created digitally in Extensible Markup Language (XML)¹. The archival descriptions can be considerable in length and numerous in numbers within a finding aid or fonds. The digital finding aids, which are digital asset repositories, are increasingly coded in the Encoded Archival Description (EAD) standard. This standard as described in [14] is the “SGML/XML based document type definition that archives, libraries, and museums are using to create, store, and distribute descriptions of their collections.” This is possible, because XML is used to create parse-able and hierarchical object models, in our case EAD, and thus facilitates the sharing of structured data across different information systems, particularly via local networks and the Internet, and also between users and information systems. EAD is maintained by the Library of

¹ http://www.w3.org/XML/
Congress (LoC) in partnership with the Society of American Archivists (SAA) [18], and is compatible with ISAD(G) [10].

The Retrieving Encoded Archival Descriptions More Effectively (README) project aims to improve archival access by developing better computational methods for finding information in digital finding aids in EAD, such that more precise or direct, and faster access to the archival material is offered. On the one hand, we hope to contribute to archival science by deploying state-of-the-art search technology developed in the Information Retrieval (IR) field to improve access to archival material, and on the other hand we are shedding new lights on IR by testing and evaluating existing search technology on real, vast and steadily increasing amounts of richly structured cultural heritage data in the form of archival finding aids.

The remainder of this paper will deal with both issues, and is setup as follow: first, we enumerate the different topics that frame our research; second, we present the baseline README system and approach; and third, we discuss the horizon beyond the baseline with more research challenges or opportunities, such as with the Encoded Archival Context (EAC) initiative.

2. RESEARCH FRAMEWORKS

Archival material and access

The importance of work processes in archival science is explained in [27]. Resulting from these work processed are for example online digital finding aids in EAD. Initiatives have been taken to facilitate the creation of the finding aids. An instance of an open-source project that deals with creating EAD files is the project Make EAD (proMEAD)\(^2\), which is a web-based native EAD editor, developed in collaboration with the National Archives of the Netherlands. Another web-based editor for EAD is ICA-Atom\(^3\) that is multi-lingual and supports multi-repository collections. Other (commercial) XML editors are also used to create digital archival descriptions in EAD, and hence advancing the 'digitization' of archival materials via digital finding aids, both online as well as offline. These editors use forms, effectively this means that creators and editors do not have to face and thus deal with the actual XML code directly.

In terms of archival access, the importance of user needs is stressed in [20], because the users eventually seek access to the online archival resources. It was argued that studying navigational features and contextual information is important, because these features better help users to understand the archives. This argument is advanced in [30], which suggest that interfaces need to provide a way to a navigational aid that supports users in providing local detail and global view of the finding aids. This suggestion emerged because it was found that the users were lost in the hierarchy, especially in the full text view. Moreover, when engaging with finding aids, users search for archival material from the bottom up and the fullest description necessary at those levels needs to be provided [25].

It is pointed out in [14] that it is in the nature of librarians and archivists to organize things in metadata such as Dublin Core, MARC and EAD. As such, there is no shortage of metadata in finding aids, but “it is a matter of finding the right hook to make them more accessible.”

Information Retrieval

A general view

Information retrieval (IR) deals with the representation, storage, organization of, and access to information items [1]. In [24] a succinct overview of the history of Information Retrieval (IR) research is given. IR research consists of two parts: automated indexing and automated retrieval. This research has been done for fifty year, and has become increasingly solid [24]. However, the impact on operational library and information systems has been slow and uneven, an area where we (and this paper) contribute to.

There is an active sub-field within IR called Focused retrieval. Focused retrieval goes further than standard IR as it tries to remove the burden on the end-user by providing more direct access to relevant information within a document [11, 23]. For lengthy and complexly structured EADs, it would save users time and effort in locating the archive they want to access.

Focused retrieval on archival finding aids

There is a range of applications within focused retrieval, such as retrieving text passages, retrieving answers to questions, and XML element retrieval by retrieving arbitrary parts of XML files. The latter is an application of focused retrieval that resembles most strongly with the approach as discussed in this paper and attempts to use the XML markup of documents to the fullest. This markup is used to represent the different levels of granularity or complexity (see Fig. 1) of possible interesting text objects. The EAD markup is mostly logical, but EAD also has document-centric features as the markup is also used for the presentation and layout. This granularity can be explicitly seen as structural hints, and used to improve the retrieval of the actual text objects.

An example is the work of [23] with XML Element retrieval on mostly scientific articles from the publisher IEEE. As archival finding aids are richly structured documents, with a complex model of information organization, finding relevant text objects in the files can be difficult. Not only because of the complexity of the organization of the archival material, but also because of the length of the archival descriptions. The quest to provide better access to EADs could be aided by technology such as XML element retrieval. Besides focused retrieval of archival material and other archival information within a finding aid, we can also contribute to improving the archival intelligence [29] of users and visitors of the archives, in other words, enhancing the

\(\text{Figure 1: EAD/XML markup and granularity: presentation, logical, semantic aspects.}\)
understanding of the archival material and the approach of working with these resources through improved usability, resulting not only in focused, but also effective access.

Importance of context
Context is a major concept for archival finding aids. The context of a finding aid partly makes content data significant and of (high) quality, besides also the form and structure. If the structure and context is detached from the actual information, then a finding aid is de-contextualized, and loses its value. Without the (logical) relationships, an archive can facilely degrade to just a collection of historical documents, or as put it in [27]:

Reliable information becomes unreliable information, high quality information degenerates to information of poorer quality; archives degenerate to documentary collections, evidence turns into documentation, documents into loose data.

Therefore, the main problem in the retrieval and presentation of content data within a finding aid is not only the actual retrieval of the desired information, but also not de-contextualizing the information at the same time. This is one of the major axioms within archival science, and one that we keep in regard. Context is also a relevant feature in IR, and can be used as a common denominator to bridge the gulf.

3. SYSTEMS AND APPROACH

Motivation
Objectives
An effective approach to focused retrieval of archival material, which could enhance archival access, is an intricate challenge. Therefore, we are addressing the following two research objectives.

1. Study effective retrieval techniques tailored to focused retrieval on archival finding aids, taking into account the user’s profile and context, the structural context, and the contextual content, of the unit to return.
2. Enhance user access to archival material through digital finding aids from multiple sources.

This paper contributes to the research conducted to fundamental approaches dealing with focused retrieval and focused presentation of archival data. We address the objectives by implementing and testing a search system that offers more focused archival access.

Requirements
Archival practices and principles. The system needs to be compliant with existing archival practices. A key archival principle is respect des fonds or the Principle of Provenance; all records of one creator are kept together. Another key principle is Respect for Original Order; all records are maintained in the order the creator had them. It is important that the autonomy of the fonds is respected.

Generalizability. The aim of this article is to give system recommendations and best practice guidelines with the README approach. Henceforth, this approach should be generalizable by other researchers and practitioners in this field as well. Moreover, we validate our approach by buckling it down to different collections from different institutions, which each have different characteristics despite using the same EAD standard.

Open-source. The software and resources that were used should be freely available. We also plan to release our tools and scripts open-source as well. It further facilitates realizable replication of our approach, making our process and results as transparent and creditable as possible. Wherever possible, we stick with state-of-the-art software that is yet to mature, but illustrate the latest (technological) possibilities. Moreover, it means our approach and achieved results can be replicated without any financial investment in software.

Overview of System Architecture
We detail the design of a state-of-the-art vertical search engine, README, for archival descriptions. An overview of the design of the README architecture is depicted in Fig. 2, in which we follow the conventional 3-tier approach of data storage, retrieval, and the eventual presentation to the user.

The README systems are developed in an out-of-the-box Fedora Core Linux operating system environment, and it is running in this environment as well. The software is also running under the Apache web server. The PC that we use is a standard desktop computer with a dual-core Intel(R) Pentium (R) processor 3.00GHz (no hyperthreads), 200Gb hard-drive, and with 2GB physical main memory.

Data
The digital asset repositories are collections of digital archival finding aids from different institutions, which also differ in length, complexity of structure, and language. The bulk of these finding aids were collected from National Archives of the Netherlands (NA), the International Institute of Social History (ISSH) located in Amsterdam (the Netherlands), and the Archives Hub (AH) in the UK. Moreover, on a smaller scale, we obtained over a hundred of finding aids from the University Libraries of
the University of Amsterdam (UBA) and the Leiden University (UBL).

<table>
<thead>
<tr>
<th>Instit.</th>
<th>Files</th>
<th>File size (bytes)</th>
<th>Lang.</th>
</tr>
</thead>
<tbody>
<tr>
<td>AH</td>
<td>3,119</td>
<td>Min 1,697, Max 889,218, Mean 9,301</td>
<td>English</td>
</tr>
<tr>
<td>IISH</td>
<td>2,866</td>
<td>Min 2,048, Max 2,922,445, Mean 35,362</td>
<td>Multi.</td>
</tr>
<tr>
<td>NA</td>
<td>2,174</td>
<td>Min 5,787, Max 10,720,767, Mean 205,749</td>
<td>Dutch</td>
</tr>
<tr>
<td>UBL</td>
<td>109</td>
<td>Min 5,577, Max 2,616,931, Mean 136,775</td>
<td>Multi.</td>
</tr>
<tr>
<td>UBA</td>
<td>60</td>
<td>Min 8,984, Max 51,677, Mean 19,196</td>
<td>Dutch</td>
</tr>
</tbody>
</table>

Table 1: General statistics of the finding aids: number of files, file size, and language.

Both libraries have adopted EAD for their special collections and have a relatively small but valuable sample of EADs. The International Institute of Social History and the Dutch National Archives are one of the few institutions in the Netherlands that have numerous full-sized and very complete EADs.

The finding aids from the NA are completely Dutch, those from the AH are completely English, and the EADs from the IISH are a mix of languages, mostly Dutch (about two-thirds of total), but for instance also German and English. Topic-wise, many finding aids from the Dutch National Archives are about Dutch government agencies, whereas the finding aids from the IISH can be related to topics about social-economic history such as archives about communists and socialists, the Archives Hub's finding aids detail the collections of libraries and museums in the UK.

The sum size of the 8000+ finding aids is 654.5 MB. The finding aids from the Dutch National Archive are significantly larger and lengthier than those from the other two institutions.

Preprocessing of the data

The data that we obtained were unverified preliminary full drafts of the archival descriptions. As a result, we had to pre-process these files in order to make them machine readable as XML. This is a prerequisite, because our approach is fully XML-driven and we can only process data that is at least well formed XML.

For instance, the finding aids from the Archives Hub were in SGML, which had to be converted to XML. Although the finding aids from the NA and the IISH were in essence XML, a considerable subset of their files was not truly well-formed XML, as some elements were not properly closed, or valid XML given the EAD specification in the Document Type Definition (DTD) or the XML Schema. Clearly, different expressions by different institutions of the EAD standard are possible, resulting in different XML code, and our approach can deal with these variations robustly. However, some uniformity such as the same set of elements as specified in the EAD standard is necessary. The uniformity is effectuated by pre-processing the files from the different institutions.

Indexing and search

Archival data encoded in EAD is structured data. Commonly used relational databases do not provide a perfect solution to store this type of data. XML databases are developed instead to provide a better solution to capture and preserve the richness of the structure in a data-structure. There are several open-source solutions available, such as eXist [17]. Other alternatives tailored specifically to archival finding aids in EAD are PLEADE (EAD on the Web) [22], Cheshire3 [15] as used by the Archives Hub in the UK, Archon developed at the University of Illinois [21], or the Digital Library eXtension Service (DLXS) software of the University of Michigan4. However, the README systems are based on another open-source solution: MonetDB [2] with the XQuery front-end Pathfinder [26] and its information retrieval implementation PF/Tijah [8].

The archival finding aids from the different institutions are indexed in a single main memory database, but in different indexes, where the 8000+ finding aids were processed and stored within minutes. The indexes are built without removal of stop words. Morphological normalization was applied on the words though by using a language-dependent stemmer for each finding aid. The document structure and order is fully preserved in this database, important information that is needed for focused retrieval of the finding aids and dealing with their context.

The queries are processed with XQuery templates. Different templates were used for each of the three README systems. Currently, we do not support yet the use of Boolean query operators (i.e. ‘and’, ‘or’, ‘not’) that is common in conventional information search systems. It is possible to do faceted search by restricting a query to a certain field like <TITLEPROPER> and selecting the collection that one wants to search exclusively in.

Ranking

A core task of IR is the matching process, i.e. given the information need of the user as expressed in a query, and a set of documents where this information can be found, what is the best (or exact) match between this query and a subset of these documents? This matching process is modeled mathematically or statistically, which is then called an information retrieval model.

The matching processes of the README systems are based on a unifying model that is called Language Modeling (LM) [19]. The essential idea in LM is that given a corpus of paired discourses, A and B, correlations can be established between the features of A and the features of B, so that for a new A, a new B can be estimated [24]. In IR, this means A is the query and B is a relevant document.

LM is an active area of research within IR and other research fields as well, because this general technique is effective for retrieval. We used the standard LM implementation of PF/Tijah as it was available and works in conjunction with our data storage component MonetDB. Using LM, we compute matching scores, which are used to rank the results in descending order according to relevance. As we work with XML files, the system returns any and arbitrary parts (depending on the focus of the granularity) of an XML file and rank these parts separately.

Presentation

Context as interface technique

The importance of context as an interface technique for making documents more understandable is discussed in [7]. Context as an interface technique for IR means that the set of found documents by a system is placed in the environment of other information types. Explicitly, context means showing the relationship of the

4 http://www.dlxs.org/
finding aids with keywords of a search, collection overviews, descriptive metadata, hyperlink structure, document structure, and the relationships to other documents within the set of finding aids. Users are getting lost in the hierarchical structure of archival finding aids [30], and to solve this problem, the idea of a user interface that could provide contextual navigation was floated. Such a presentation would support users by providing both the local detail and a global view of the relevant information. Ideally, this would make archival finding aids no longer barriers, but more boundary spanners. It is important to show relevant information in context [3, 16]. The findings in [12], where a study was conducted using a scientific collection of documents (not EAD), also suggest that users appreciate presenting information in context more.

**Document Order-Structure-Depth Model**

The presentation of focused retrieval of archival material remains an open question. That is why we propose in Fig. 3 our Document Order-Structure-Depth (DOSD) model, which captures our assumptions comprehensively. We use this model as a principle to present and display each retrieved result from an EAD/XML file in context in a user interface, given the document order, the structure and the depth.

A (part of the) screen can be represented as a Cartesian plane, with the X-axis the depth, and on the Y-axis the structure (granularity, complexity) of the fonds. For example, retrieved text objects that appear in the second quadrant have little depth, little structure, and are in the top of the archival finding aid. Our supposition is that this model could intuitively give focused access to archival material in a natural way. More future research is needed to effectively discover the potential merits and inadequacies of this model.

**Hitlist in context**

The hitlist is the list that is returned by a system with ranked results; after the user has entered the query, and the system has computed the matching scores given the query and the EAD files. Since this is the first display that the user sees after entering the query, and the first stage of assessing the relevancy of the results, it is worthwhile to investigate not only what is returned, but also how and why. We believe we can provide more focused access to the archival material by showing relevant results directly (in

![Figure 3: Document Order-Structure-Depth (DOSD) Model.](image-url)

**Figure 4: Archival Material in Context (AMC).**
context); providing access to only the (beginning of an) entire fonds is therefore neither immediately necessary nor desired. We materialized the Archival Material in Context (AMC) system as depicted in Fig. 4 with that idea in mind. It is an implementation of the DOSD model as discussed before. We used the query “juliana greet hofmans”, with the intention to search for information related to former Dutch queen Juliana (1909-2004), her adviser Greet Hofmans (1894-1968), and the subsequent crisis in Netherlands in the 50’s of the 20th century. We use this query as an example for all the three systems.

As reported in [32] of presenting a focused hitlist in context, namely preserving provenance by grouping most relevant individual items together per finding aid (and thus creator); preserving document structure and returning the individual archival items in the hierarchical document order, such that the local and global context of a finding aid can be combined and the archival bond of a fonds is kept in regard; and finally allowing deep-linking and direct access so that the user can get actual focused access to the individual items by optimally exploiting the full context. Individual results can be put in context given the hierarchical XML tree by either showing its ancestors or descendents. The latter is however not always really usable from an IR point of view, because any information in the descendents is already known in the current node which results in overlap of information.

Alternative hitlists
Besides the AMC system, we developed two alternative versions (see Fig. 5 and 6) that retrieves and provides access to the finding aids on a different granularity level, namely on the file level (only top) and level element (anything between top and bottom).

Whole Fonds (WF)
The Whole Fonds system as shown in Fig. 5 ranks and retrieves an entire finding aid (document), and is comparable to a conventional document retrieval system like Google or Yahoo. For each result, a title and a snippet (short preview of fonds) are presented.

Individual Archival Material (IAM)
Fig. 6 shows the Individual Archival Material system, that retrieves XML element nodes as natural units, and it is therefore comparable to a standa

Alternative hitlists
Besides the AMC system, we developed two alternative versions (see Fig. 5 and 6) that retrieves and provides access to the finding aids on a different granularity level, namely on the file level (only top) and element level (anything between top and bottom).

Whole Fonds (WF)
The Whole Fonds system as shown in Fig. 5 ranks and retrieves an entire finding aid (document), and is comparable to a conventional document retrieval system like Google or Yahoo. For each result, a title and a snippet (short preview of fonds) are presented.

Individual Archival Material (IAM)
Fig. 6 shows the Individual Archival Material system, that retrieves XML element nodes as natural units, and it is therefore comparable to a standard XML element retrieval system that retrieves arbitrary parts of a XML document. Besides the title and the snippet of the element, we also show its result path in XPath.

Assessing systems in user study
In [4] we conducted a user study to assess the README system as outlined here. An empirical study was conducted with 9 test persons with sessions that lasted 1.5 hour on average for each participant. The AMC system was compared against a system that would return whole fonds (WF), and one that only returns the
individual archival materials (IAM). In both systems, the context is omitted, and using this comparison we can examine empirically the effects of the context in the hitlist. The experiment consisted of a series of questionnaires with random iterations of interaction with the three systems. Table 2 shows post-task questions and the responses toward features in the three different types of hitlists.

Q3.13 How satisfied were you with the information provided in the hitlist?
Q3.14: Was the overview of results clear?
Q3.15: Was it easy to select the most promising result?

The overview of the results was found most clearly in the WF system (Q3.13), likely because of its simplicity and it is conventional (and thus familiar) presentation. Henceforth, the test persons tend to be most satisfied with the information provided on the hitlist of the WF system (Q3.14). However, they found it easiest to select the most promising result in the AMC system (Q3.15). The IAM system was least appreciated. The results of the user study show that AMC system is not optimal, but achieves its objective of offering users focused archival access. The study gave concrete suggestions on how to improve the user interface by presenting the context in a more intuitive way, which we will explore in future research. Effectively, it means combining the best of the WF and AMC interfaces.

4. Concluding Discussion: Beyond Baseline

This paper focused on the question of effective access methods, by developing novel search tools that will be crucial on the massive scale of digital asset repositories. We illustrated concretely why XML matters in digital curation by presenting a fully XML-driven system description for digital assets. Some of the challenges that we faced to improve information access in the archives were identified. We proposed an approach to deal with these challenges.

However, there are still roadblocks lying ahead in terms of providing information access with EAD. For example, the ranking of the results, especially on the element level, has not been optimized yet in the IR model – crucial in providing focused access. To optimize the ranking of the results, we will conduct experiments to discover optimal settings in our retrieval models for retrieving desired archival descriptions more effectively - at least the ones that are available to our research by creating an EAD test collection.

The research in this paper has been centered on the retrieval and presentation of the archival descriptions from a document-centric and hierarchical structural point of view. Intrinsically, other views exist with additional applications of XML. For example, a promising direction is to help enrich EADs with link detection methods, and provide access to the archival descriptions by exploiting additional relational structures besides the hierarchical structure, which we have done so far. In other words, certain texts in a finding aid can be clicked and directs a user to a different finding aid or a different point in the same finding aid. There could be special use for automatically generated links within a

fonds itself, specifically the result display as illustrated in Section 3.7.5. In case the user chooses to go beyond the hitlist, usually in the case of serendipitous information seeking (‘browsing’) task, then EADs enriched with links could provide additional focused access to the archival material by saving the user browsing time.

In [33] we set the first steps in this direction by presenting preliminary work on this topic, where we showed we could automatically detect occurrences of person names with high accuracy, both in and between archival descriptions. This allows us to create (pseudo) encoded archival context descriptions that provide novel means of navigation, improving access to the vast amounts of archival data not only through the inventories, but also through the actors. This means that besides discovering relationships between the fonds in one collection, we can also detect them between the fonds in the same collection, and even between different institutions. The concept of parallel provenance is strongly related to this, and is addressed by Ketelaar [13], which he paraphrased as “two or more entities residing in a different context as establishing the provenance of a record, even when they are involved in different kinds of action, for example creation and control.”

Archival context may be constructed through the use of authority records that capture information about the record creators or actors (corporations, persons, or families) and the context of the record creation. By separating the record creator’s descriptions from the records or resources descriptions themselves, we can automatically create ‘links’ from all occurrences of the creators to this context. The resulting descriptions of record creators can be encoded in XML and matched using the emerging Encoded Archival Context (EAC) standard.

Currently, EAC has only been applied experimentally. One of the main barriers to adoption is that it requires substantial effort to adopt EAC. The information for the creator’s authority record is usually available in some form (for example, EAD descriptions usually have a detailed field <BIOGHIST> about the archive’s creator). However, linking such a context description to occurrences of the creator in the archival descriptions requires more explicit structure than that is available in legacy data.

Having established these relations, we can create physical links by directly linking two or more fonds together, for example in XLink. We can also extract information existing in another fonds to create pseudo archival context descriptions, or we can even automatically construct an authority record in EAC by discovering co-references. These are all steps towards even more effective information access using EAD.

5. ACKNOWLEDGMENTS

Junte Zhang gratefully acknowledges Henny van Schie of the National Archives of the Netherlands for his support and insightful discussions on EAD. This research was supported by the Netherlands Organization for Scientific Research (NWO) Innovative Research Grants Program (VIDI Scheme) under project number 639.072.601.

6. REFERENCES


Integrating Metadata into the NARA Transcontinental Persistent Archive Prototype via the OAI-PMH

Jewel Ward, Jennifer Mantooth
School of Info. & Library Science
Univ. of North Carolina at Chapel Hill
Chapel Hill, NC 27599 USA
jewel_ward@unc.edu, jnmantoo@email.unc.edu

Antoine de Torcy
National Center for Data Intensive Cyber Environments (NC-DICE)
School of Info. & Library Science
Univ. of North Carolina at Chapel Hill
Chapel Hill, NC 27599 USA
antoine@diceresearch.org

Mason Chua, Jonathan Crabtree
H.W. Odum Institute for Research in Social Science
Univ. of North Carolina at Chapel Hill
Chapel Hill, NC 27599 USA
{aftran, jonathan_crabtree}@unc.edu

ABSTRACT
The H.W. Odum Institute for Research in Social Science (Odum), the Renaissance Computing Institute (RENCI), the School of Information and Library Science (SILS), and the National Center for Data Intensive Cyber Environments (NC-DICE), all part of the University of North Carolina at Chapel Hill (UNC-CH), are collaborating on an extension of the National Archives and Records Administration's (NARA) transcontinental persistent archive prototype (TPAP) data grid with the new integrated Rule Oriented Data System (iRODS). The goal of the project is to enable collection interoperability among preservation environments around the TPAP data grid using iRODS. This paper presents the results of one part of that project, which is the development of a prototype service by which metadata can be transferred into the NARA TPAP metadata catalogue (iCAT) via the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH), using the Odum Institute Data Archive’s (OIDA) Data Document Initiative (DDI) metadata as the test data. The successful development of this prototype will enable bibliographic metadata-based queries in iRODS, as well as enable any digital library or data archive that is an OAI-PMH-compliant Data Provider (DP) to upload their metadata into the preservation grid. Future work will include ingesting the digital objects, as well.

Categories and Subject Descriptors
H.3.4 [Information Storage and Retrieval]: Systems and Software – distributed systems, information networks

General Terms
Standardization.

Keywords
iRODS, NARA TPAP, OAI-PMH, DDI, metadata, digital

This work is licensed under the Creative Commons Attribution-Noncommercial-No Derivative Works 3.0 Unported license. You are free to share this work (copy, distribute and transmit) under the following conditions: attribution, non-commercial, and no derivative works. To view a copy of this license, visit http://creativecommons.org/licenses/by-nc-nd/3.0/.

DigCCurr2009, April 1-3, 2009, Chapel Hill, NC, USA
OAI-PMH-compliant Data Provider (DP) will be able to upload their metadata into the preservation grid. A future version of this prototype will include ingestion of the objects themselves. The ability to seamlessly migrate content from diverse digital content systems into the preservation grid will minimize risk of data loss as well as enhance versatility. We envision this capability as useful not only to preserve collections within the grid environment, but also to empower users with the ability to migrate complete archives out to other digital archive platforms. This migration approach will help insulate from the risks associated with technology turnover and could potentially be another tool for preservationists.

2. BACKGROUND
The Transcontinental Persistent Archives Prototype is the result of more than eight years of ongoing research based on collaborations that include the University of California at San Diego, the University of Maryland, the Georgia Institute of Technology, the National Archives and Records Administration, the Alleghany Ballistics Laboratory and the University of North Carolina at Chapel Hill. NARA’s mandate is to preserve data for “the life of the Republic”. The purpose of the TPAP project is to provide a test bed with which NARA can evaluate methods for preserving electronic data [6].

This data preservation environment uses Storage Resource Broker (SRB) and iRODS data grids. SRB is middleware that supports organization of distributed data into shared collections across heterogeneous storage systems and multiple organizations. The software infrastructure has “features to support the management, collaboration, controlled sharing, publication, replication, transfer, and preservation of distributed data” [7].

The developers of iRODS based its design on this proven SRB data grid technology and extended it to give administrators the ability to automate individualized community-based management policies [8]. There are six nodes in the TPAP data grid, each of which manages its own environment, metadata catalog, and storage system [9]. These federated nodes are located in Washington, D.C., West Virginia, California, Maryland, and North Carolina. This synchronized federation of data enables preservation by replicating the data across nodes so that there is no one point of failure.

A group of digital librarians and researchers created the OAI-PMH in 1999 as a way for digital librarians to provide access to e-prints across multiple repositories [10]. By implementing the protocol on top of an existing digital library while at the same time providing catalog information about materials stored elsewhere, digital library administrators could make it easier for users to find the resources they need. By 2002, one hundred digital library administrators had registered their repositories with the OAI-PMH, thus providing access to just over 1 million records [11]. By 2006, the number of registered repositories had grown to 776 and they contained over 10 million metadata records [12]. As of January 2009, digital library administrators have registered 943 repositories; the number of additional metadata records that can be added to the previous total has not yet been counted [13].

The continued growth of the OAI-PMH after almost 10 years has shown that it is the de facto standard for metadata transfer within the digital library community. The researchers involved with the eight-year old NARA TPAP project use SRB, a proven technology, and iRODS, a second generation data grid. As digital librarians and researchers have turned their attention to the preservation of the collections contained within the digital libraries, and as the size and number of these collections continues to grow, the attention of these administrators has focused on using existing data grid technologies such as iRODS to preserve their data. In order to use iRODS, digital library administrators will have to agree on a method with which to ingest those collections into the preservation grid.

As we considered the best way to ingest metadata and digital objects into the NARA TPAP from the OIDA, we decided to examine whether or not we could leverage existing, established, commonly used and proven technologies for this purpose. We decided to evaluate whether or not we could use the OAI-PMH to transfer metadata from the OIDA into the iRODS iCAT via a microservice. Microservices are small, well-defined procedures/functions written in C by systems and application programmers that perform a specific task when compiled into iRODS server code [14]. We began the evaluation by creating a metadata extraction and ingest prototype.

3. TECHNIQUES AND METHODS
The metadata extraction and ingest prototype works as follows:

1. A DDI XML file from the OIDA’s OAI-PMH DP is obtained as the source file.
2. The DDI XML file is validated and ingested into iRODS as an object.
3. A microservice creates an HTML file that contains the URLs to digital objects pointed to in the original DDI XML file.
4. The DDI XML file is formatted via an XSL Transformation. A microservice parses the bibliographic metadata into an Attribute-Value-Units triplet, stores each metadata triplet in the iRODS iCAT metadata catalog, and associates the metadata with the HTML file in (3) above. A metadata Attribute-Value-Units triplet (AVU) consists of an Attribute-Name, Attribute-Value, and one or more optional Attribute-Units [15].
5. A user can now access the DDI XML file via a web interface to iRODS, via the search field. A search on the terms listed in the AVU triplets will retrieve the HTML file. The user opens the HTML file, clicks on a link, and obtains access to the desired digital object.

4. RESULTS AND DISCUSSION
The prototype thus far has been a success as a proof-of-concept. We have built on existing, validated standards and systems such as iRODS and the OAI-PMH. The current implementation of the prototype moves the DDI XML file into iRODS as an object, while 22 metadata elements of the DDI XML file are moved into the iCAT AVUs. This prototype has enabled bibliographic metadata-based queries in the iRODS iCAT.
Initially, we planned to enter all metadata into the iCAT AVUs without saving the original DDI XML files, but we later reconsidered and decided to upload the DDI XML file as an object. We used an XSLT to ingest the DDI elements into the iCAT AVUs. We chose the 22 DDI elements based on existing search fields in the Odum Institute Digital Archive such as title, author, global ID (handle), abstract, and keyword, etc.

Our decision to move the entire DDI XML file into iRODS as an object means that a user can download the original DDI XML file for ingestion into their data archive without having to “reverse engineer” the file from the iCAT AVUs. iRODS is both an archival and dissemination system, but it can also be a transfer system, where digital libraries both pull and push digital objects and metadata.

One benefit of using stylesheets to transform the metadata in conjunction with microservices is that different stylesheets may be developed to serve different needs and audiences. One example of this would be to accommodate users with different levels of access to the data and metadata.

5. CONCLUSIONS AND FUTURE WORK
The outcome of this prototype is the proof-of-concept that creating a large-scale metadata and digital object extraction and ingest tool for batch upload to the NARA TPAP is achievable. A seamless batch upload process would be both practical and useful for archivists and digital archive managers.

Our next steps will involve determining users’ search habits by analyzing logs of their search history before further refining the XSLTs. The overall model needs to be extended so that the DDI XML files can be transferred and transformed as batch uploads. A needs assessment will be conducted to determine whether or not additional preservation metadata should be added at the point of ingestion into iRODS. As well, we must create a service to verify that the metadata transferred from the digital archive is actually present in the iCAT.

A major area of research is to implement a process by which the digital objects pointed to by the URL links in the DDI file are ingested into iRODS. From a digital curation perspective, it is preferable to store the metadata and all digital objects that it references within one preservation environment, rather than rely on external systems in which persistence is not guaranteed. We believe that the future ability to migrate between digital preservation environments will require that both the objects and their associated metadata reside within the preservation grid.

6. ACKNOWLEDGMENTS
We would like to thank our reviewers for their valuable comments. This work is funded by the NSF grant OCI-0848296 and is a collaboration with NARA on the development of the "NARA Transcontinental Persistent Archive Prototype". The initial work on this project was funded by the NARA supplement to NSF SCI 0438741, “Cyberinfrastructure; from Vision to Reality” – Transcontinental Persistent Archive Prototype (TPAP) (2005-2008).

7. REFERENCES
Creating Metadata for a Digital Database: A Case Study

Joanne She
DOJ Legislative History Digitization Project
U.S. Department of Justice Libraries
Washington DC 20530, USA
Joanne.she@usdoj.gov

Myron B. Chace
Preservation Reformatting Division
Library of Congress
Washington, DC 20540, USA
mchace@loc.gov

ABSTRACT
Digitizing a complex collection that contains a variety of content types creates a great challenge to make the entire contents of the documents fully accessible to meet researchers’ needs. The authors provide a case study that describes an innovative approach to create metadata for a complex legislative history digital database with several features: (1) Focus on users’ legislative history research needs to select search fields beyond the traditional access points; (2) Innovatively apply the Dublin Core Metadata Standards and the concept of Functional Requirements for Bibliographic Records (FRBR) to contextually design the database and user interface; (3) Contextually create metadata for legislative documents in order meet complex research needs: locating all the documents of a particular legislative history; finding related documents, or a particular document; searching a particular author/sponsor’s information; searching multiple legislative histories. The authors also provide sample metadata records, sample user interface screen with search features. This case study shows that digital technology helps not only to convert information from a variety of source materials to a single accessible (digital) format, but also offers multiple access points to match complex research needs because digital technology can contextually organize information in an active form.

Categories and Subject Descriptors
D.2.2 [Design Tools and Techniques]: Modules and interfaces, Structured programming, User interfaces; D.2.10 [Design]: Methodologies; D.4.3 [File Systems Management]: Access methods, Directory structures, File organization; D.4.7 [Organization and Design]: Hierarchical design, Interactive systems; H.3 [Information Storage and Retrieval]: Indexing method; H.3.3 [Information Search and Retrieval]: Query formulation, Retrieval models, Search process, Selection process; H.3.7 [Digital Libraries]: Collection, Dissemination, Standards, User issues; H.5.2 [User Interfaces]: Screen design, Theory and methods, User-centered design

Keywords
Digitization, knowledge management, metadata creation, legislative history research, legal research

INTRODUCTION
Research in legislative history is important in United States public law study, since most legal decisions and rulings are based on interpretations that become precedents. As Robert Berring mentioned in his book Finding the Law: “One way of determining what a statute ‘really means’ is to perform a legislative history [1].” The complicated U.S. legal system, however, has made legislative history research more difficult because such research often encounters a series of steps in a complex legal process. “A traditional legislative history is the gathering together of every relevant document that was part of the process of enacting the statute [2].” “Legislative histories can be very extensive, especially for laws with many titles and sections and on legislative matters that may have been percolating over several congresses in various legislative measures [3].” Thus, a legislative history is a complex compilation of relevant legislative documents that provide the history of a law. What do researchers want to find out in legislative history research? “The belief is that by examining the evidence of what the legislators knew and said at the time it acted, the researcher can discover what they really meant. In the matter of federal legislation, this process leads one through a large number of documents of various types [4].” Even though legislative history research is important but complicated, it seems that there was no efficient reference tool besides the Congressional Index Services.

Generally researchers first have to find the Popular Law Name and Public Law Number, then locate a document list from the CIS Index, then find the document based on the citation or the access number. Sometimes, there is no identification such as a “doc number” or CIS number. Researchers often have to devote extended periods of time to find the documents one-by-one through a great number of books, microforms, or databases. Finally, researchers can pull the documents together, but only after this long, sophisticated search process. In addition, the CIS Index has limited time coverage, since it covers legislative history from 1970. Annual editions usually lag two years. These constraints increase difficulty for researchers to do legislative history research on older as well as current laws.
GREAT CHALLENGES

Emerging digital technology offers an opportunity to digitize complex legislative history documents in order to make them more accessible. Scanning and converting scanned documents into searchable pdf files initially appeared not to be difficult. However, making legislative documents fully accessible contextually is a great challenge absent a knowledge management tool that could be used directly to organize complex digital legislative documents. A review of “Dublin Core Metadata Standards” revealed that most elements apparently are not useful for describing legislative documents.

The Dublin Core User Guide employs a “one-one principle” which means only a single manifestation or version of a resource can be described. With such a limitation, how can complex relationships in legislative documents be captured in metadata?

Consulting standard works about bibliographic control such as Anglo-American Cataloguing Rules (AACR2), OCLC Bibliographic Formats and Standards, and Cataloging Legal Literature also provided no significant help. Again, the treatment of a legislative history (compilation) in these works is simply one book/publication without further descriptive information about contextual relationships within that history.
INNOVATIVE DESIGN

Designing a metadata template that suits complex legislative history research requires: (1) Selecting appropriate access points and (2) Contextually structuring and arranging legislative documents.

(1) Access points. Traditional access points are not sufficient for research in legislative history.

Figure 7: Traditional access points (from the Online Catalog of Library of Congress.)

Can there be additional and modified search fields which better suit legislative research? Based on the User Guide: Using Dublin Core: “Each element is optional and may be repeated,” plus the basic set can be extensible meaning the search fields are modifiable. This flexibility made selection easier. Focusing on legislative history research needs, 12 core elements were selected as search fields: Popular Name, Public Law Number, Title, Author/Sponsor, Bill Number, Report Number, Document Number, Citation, Congress, Date, Publication Type and Subject. Some elements are further subdivided. For example, Publication Type is subdivided by: Act (Public Law), Bill, Report, Debates, Presidential Statements. And usually Bills, Reports, and Hearings are subdivided by House and Senate.

Figure 8: 12 core search fields:

2) Context structure and legislative document arrangement. Necessary for legislative research is finding a contextual legislative history, which “is a more active form of research that will move beyond the normal federal materials. This approach can help one to focus on certain documents, or even parts of documents, making the whole process simpler [5].” A basic research need, evidently, is to find contextual legislative history information in an active form instead of a traditional way.

Are there steps to make Dublin Core’s “The One-to-One Principle [6]” (which in general allows only one description of a manifestation or version of a resource) work to describe the complex relationships among legislative documents? Based on underlying concepts in The Bibliographic Record and Information Technology: “Access implies a mechanism for isolating a single record or related group of them from among many others... Both processes involve 1) Selecting one or more useful access points for each document from among its data elements. 2) Structuring and arranging these access points... 3) Linking each access point... 4) Linking related access points with each other [7].” Thus access points identified above could be arranged and linked with each other. But how? Contextual arrangement and linking should be based on relationships according to the concepts of Functional Requirements for Bibliographic Records (FRBR): “In the context of the model, relationships serve as the vehicle for depicting the link between one entity and another, and thus as the means of assisting the user to ‘navigate’ the universe that is represented in a bibliography, catalogue, or bibliographic database. Typically the user will formulate a search query using one or more attributes of the entity for which he or she is searching, and it is through the attribute that the user finds the entity sought. The relationships reflected in the bibliographic record provide additional information that assists the user in making connections between the entity found and other entities that are related to that entity. Relationships may be reflected ... in a number of ways [8].” How can all the legislative history documents be linked together based on their complex relationships? By examining the law-making process and its relationships, one can see the entire law-making process actually is the process of “how a bill becomes law.” The central element is the bill, and many documents are related to it such as hearings, debates, reports, etc. Occasionally, one law may be involved with several bills and their relationships could be more complicated.
Since all documents of a particular law are part of its legislative history, the popular law name and law number should serve as a single thread that ties every document together.

Once research needs and the relationship among the legislative documents are clarified, metadata creation becomes easier. Each document should be linked with the law being researched and related document(s). Linkage could be complicated, however, due to a complex relationship among the documents. In this situation, analyze the document(s) and select key elements from each document as determined by research needs such as citation, document number, bill number, report number, etc. A few examples follow: Figure 9-10: Act (Public Law).

Figures 11-12: Bill.

The bill number is an important key element besides the popular name and law number, since a bill is the object of debates, hearings, reports, etc. Numerous documents could be produced during the process of a bill-becoming-a-law. Here, bill-related documents can be kept together by using the central point-bill number. The date, document number and the publication type help to identify this document.

Figures 13-14: Debates.

From the metadata, one can see that this document (an act) is linked by Public Law number and Popular Name. But the act also has relationships with several bills and reports thus it can be found while searching the popular name or law number. The act (document) also can be located while searching those related documents or search by citations.
Debates are frequently related to a bill or multiple bills. In this instance, the related bill is one key element or linking point between debates and the bill. A sponsor is also an important key element for a bill and debates, because the researchers usually want to find out “who” did “what” at the debates.

Figures 15-16: Hearings.

Figures 19-20: Documents
A presidential document is usually officially published in the *Weekly Presidential Document*. Using that citation is one of the required elements for metadata.

By examining all these examples above, one can see that the Popular Name and Public Law number are two required fields besides actual title and date for each document as a main thread which ties up all the legislative documents within a law’s history. There are also multiple links among the related documents based on their relationship. Each field is repeatable, and allowed various forms.

**BENEFITS**

This innovative metadata creation makes research more active and efficient, so that researchers can easily find what they need by a simple search or browse. Here are the major benefits: Locating a particular legislative history; finding related documents or a particular document; searching an author/sponsor’s information; searching across multiple histories; conducting a combined (modified) search; modifying search within the result; using multiple ways to find the same information. In addition, all the information can be sorted based on research needs. It can be also used to track bills and to compile a preliminary legislative history on the fly. Please see the following examples:

1. **Locating a particular history** (a compilation of legislative documents: a) by search. Figure 23
   
   ![Figure 23](image)

   Figures 21-22: Presidential documents

2. **Finding related documents.** Figure 25.
   
   ![Figure 25](image)
(3) Search a particular document:
   a) by citation. Figure 26.

   Search an individual document (by citation)

   b) by document number. Figure 27.

   Search an individual document (by document number)

(4) Search across multiple histories. Figure 28.


   Combined Search (1)

   The combined/modified search could help narrow the search results.

   Combined Search (2)

   All the information is in an active form, and it can be dynamically sorted based on a researcher's needs - e.g., chronological order, type of publication, etc.

   Figure 30: search result sorted in a chronological order.
From this chronological sorting order, “what happened” and “when” can be easily determined.

b) Figure 31: Search result sorted by type of publication first, then date.

The examples above demonstrate this innovation can contextually organize legislative documents in an active form enabling researchers to find out “When?” “Who?” did “What?” during the legislative process.

(7) Metadata can also be created on the fly for current legislative documents. The same methods can be used to meta-tag current legislative documents for current bill tracking and compiling a preliminary legislative history, since digital documents are easy to add to the database.

CONCLUSION
This case study reveals the benefit of creating metadata for an innovative legislative history digital database that makes legislative documents fully searchable in an active form to satisfy complex legislative history research in addition to preserving the documents. The success of this innovation actually results from fully understanding legislative history research needs, creatively applying Dublin Core Standards, the basic theory of bibliographic control, and the concept of FRBR. Digital technology in this application demonstrates that it can be used innovatively to provide access to many sources and to fully meet complex research needs.

ACKNOWLEDGEMENT
Special thanks to: Dessy Blane, Director of U.S. DOJ Libraries, Ms. Roberta Shaffer, Executive Director of FLICC, and Dr. Ingrid Hsieh-Yee for all the opportunities and training in legal information and digital project management.

REFERENCES
Curation of Scientific Datasets: Trends, Current Initiatives and Solutions

Malte Dreyer, Heike Neuroth
Max Planck Digital Library (MPDL), Munich, Germany
{malte.dreyer,heike.neuroth}@mpdl.mpg.de

Sarah Carrier, Jane Greenberg
School of Information & Library Science, University of North Carolina at Chapel Hill Chapel Hill, NC 27599 USA
{scarrier,janeg}@email.unc.edu

Stephen Abrams, Patricia Cruse, John Kunze
California Digital Lib. Univ. of California Oakland, CA 94612 USA
{stephen.abrams, Patricia.Cruse,jak}@ucop.edu

Michael Day, Colin Neilson, Alexander Ball, Rosemary Russell
UKOLN University of Bath Bath, UK
{m.day,c.neilson,a.ball,r.russell}@ukoln.ac.uk

ABSTRACT
E-Science and cyberinfrastructure developments present information professionals and researchers with significant curation challenges relating to the management of scientific datasets [1]. Among pressing questions are: What data should be collected for data curation? How can quality control be maintained? And, how can metadata be generated effectively? These and other challenges are made complex, given the diversity of methods by which data are produced, their heterogeneity, and the increasing scale and scope of scientific research projects. Available literature on the topic of data stewardship provides grounding for approaches addressing these problems, yet more work specifically relating to cyberinfrastructure and repository frameworks is required [2]. This international panel will report on current initiatives addressing the management of scientific data, focusing on advances and solutions in the curation of datasets. The reporting will take place in the context of recommendations from funding agencies and international councils [3,4,5], and models for data curation such as the DCC Curation Lifecycle Model [6]. The panel will provide recommendations for the scope and form of the effort required to address the challenge of scientific data curation and the implications for digital curation education.

Categories and Subject Descriptors
H.3.5 [Online Information Services]: Data sharing, Web-based services; H.3.7 [Digital Libraries]: System issues, Standards

Keywords
Curation, scientific data, cyberinfrastructure, education

This work is licensed under the Creative Commons Attribution-Noncommercial-No Derivative Works 3.0 Unported license. You are free to share this work (copy, distribute and transmit) under the following conditions: attribution, non-commercial, and no derivative works. To view a copy of this license, visit http://creativecommons.org/licenses/by-nc-nd/3.0/.

DigCCurr2009, April 1-3, 2009, Chapel Hill, NC, USA

1. PRESENTATIONS
1.1 Research Data Curation: Problems and Challenges (Malte Dreyer, Heike Neuroth)
A chief motivation for data curation is discovery and re-use of valuable research data. Research universities and large organizations such as the Max Planck Institute face curation challenges due to the diversity and expanse of data produced. Specifically, they need to address issues of what data should be collected, quality control, curatorial responsibility, trust, and sustainability. An alliance of scientific organizations in Germany has been formed to collectively address these problems. The alliance includes the Deutsche Forschungsgemeinschaft (DFG, the German Research Foundation), the Fraunhofer Society, the Helmholtz Association of German Research Centres, and the Max Planck Society. All of the members have signed a joint national e-infrastructure policy initiative that has six priority areas; one is focusing on “Preservation and re-use of primary research data” [7]. An emphasis of our work is on open data. This presentation will give an overview of ongoing discussions in Germany underlying the alliance, issues and decisions made specific to data curation, and steps to support open access.

1.2 The Dryad Repository Application Profile: Groundwork Towards a Metadata Scheme for Scientific Data (Sarah Carrier, Jane Greenberg)
The Dryad Repository hosts datasets underlying papers published in the field of evolutionary biology and related sciences. Dryad’s metadata architecture links data object metadata with publication metadata. The repository metadata team has developed an application profile with functional requirements that include long-term preservation of datasets, object retrieval and reuse, versioning, provenance tracking, instantiations, and the representation of complex relationships between datasets. Dryad’s application profile supports the entire life cycle of a data object, starting with its generation, and ensures the long-term preservation of the metadata itself. The application profile is in compliance the Singapore Framework for application profiles, a framework compatible with the Dublin Core Abstract Model.
A Micro-Services Approach to Data Curation (Stephen Abrams, Patricia Cruse, John Kunze)

Data curation is a set of activities aimed at maintaining a balance of usability and authenticity of data objects over time. Rather than centering these activities around a preservation repository, we see them spread across a range of access repositories. Relatively quiescent, or even “dark,” storage systems are still important tools, but selected curation and preservation services can and should be applied to any repositories with sufficiently highly valued data assets. It follows that such services are inherently not repository-bound. For example, a naming micro-service could supply preservation-ready identifiers for newly born data objects originating in any number of laboratories within an institution or a discipline; an identity micro-service could then host the basic metadata bindings to give descriptive reality to the named object. Among those bindings, deliberately curated data should generate technical metadata as a matter of course during processing first by a characterization micro-service, to supply early feedback on well-formedness, and second by a fixity micro-service, to generate checksums to help in change detection and version management. This presentation will review key components of the micro-services approach to data curation and note some of our current challenges. We will also comment on the impact of this topic on data curation education and preparing professionals.

Disciplinary and Institutional Perspectives on Digital Curation (Michael Day, Colin Neilson, Alexander Ball, Rosemary Russell)

Abstract models like the DCC Digital Curation Lifecycle embody the concept that the curation of research data cannot be considered in isolation from the wider contexts of scientific research and practice. One aspect of this is the need for curators to engage with the teams and individuals that are responsible for creating data. Some recent studies [5, 8] have begun to identify how curation roles and responsibilities are shared across all of those institutions and individuals that play an active part in the ongoing stewardship of research data, including scientists, institutions, data centres, funding bodies, and the users of third-party data. This sharing of responsibilities for curation emphasises the importance of collaboration, and the need for generic technical and organisational frameworks to support it. In practice, however, the data curation cultures of different research disciplines (and sub-disciplines) are extremely diverse, posing significant challenges for those trying to develop generic (or institution-based) solutions. This presentation will explore these issues with reference to detailed disciplinary case-studies of curation undertaken for the Digital Curation Centre as part of the DCC SCARP (Sharing Curation and Re-use Preservation) project and a feasibility study conducted by UKOLN into the potential for developing a generic metadata application profile for scientific datasets.

REFERENCES

Technology Learning for Digital Curators

Peter Botticelli
School of Information Resources & Library Science
University of Arizona
Tucson, AZ 85719 USA
pkb@email.arizona.edu

Jana Bradley
School of Information Resources & Library Science
University of Arizona
Tucson, AZ 85719 USA
janabrad@email.arizona.edu

Bruce Fulton
School of Information Resources & Library Science
University of Arizona
Tucson, AZ 85719 USA
bfulton@email.arizona.edu

ABSTRACT
This panel discussion will explore the current technology learning needs of information professionals who are involved in curating digital collections. We will examine the problems faced by curators from a number of perspectives, including professional communities, institutions that own or manage large scale collections, and current students who are now developing careers in digital curation. We will also discuss an ongoing research effort that explores the use of experiential or hands-on methods for authentic learning of the technologies and work functions involved in digital curation.

Our panelists were chosen to represent a variety of institutions and professional communities involved in curation-related functions or activities, and having a strong interest in the development of advanced technology skills among information professionals.

Categories and Subject Descriptors
K.3.1 [Computing Milieux]: Computer Uses in Education – Distance learning; K.3.2 [Computing Milieux]: Computer and Information Science Education – Curriculum, Information systems education, Literacy

Keywords
Digital curation, experiential learning, virtual laboratories

PANEL MEMBERS
Peter Botticelli is Assistant Professor of Practice in the School of Information Resources and Library Science (SIRLS) at the University of Arizona. He coordinates the Digital Information Management (DigIn) certificate program at SIRLS, and also teaches courses in the certificate program as well as in the SIRLS master’s program. He is directly involved in the School’s current research on virtual laboratory exercises for curation.

Bruce Fulton, Digital Projects Librarian at SIRLS, serves as the project manager and principal technology liaison for the IMLS-funded research project entitled: “Improving Student Learning of Advanced Digital Technologies in an Online Laboratory: A Research Approach.” Fulton teaches courses both for DigIn and the SIRLS master's program, and has developed all the hands-on laboratory and virtualization components used in the DigIn program to date.

Richard Pearce-Moses is Deputy Director for Technology and Information Resources, Arizona State Library, Archives and Public Records, and is a former President of the Society of American Archivists. Pearce-Moses is an internationally recognized advocate for the development of digital curation skills in the information professions, particularly archives and records management. He also played a key role in the founding of the DigIn program and has been a regular advisor to the program.

Tyler O. Walters is Associate Director for Technology and Resource Services, Georgia Institute of Technology Library and Information Center. Walters’s perspective on digital curation is informed by his leadership in digital preservation efforts as well as the building of institutional repositories and digital publishing services. In spring 2009, Walters will be teaching a course in the DigIn program, entitled “Managing the Digital Information Environment.”

The panel also includes three current students in the DigIn program:

Alvin Hutchinson, Information Services Librarian, Smithsonian Institution Libraries, manages a digital repository of scholarly research publications and serves on the Community Outreach Group organized by the DSpace Federation. He is responsible for all aspects of the Smithsonian Digital Repository including policy, content management and technological infrastructure.

George Diez, Librarian, National Library of Education, is currently involved in managing digital resources and planning digitization efforts. He is particularly interested in open access issues, the building of institutional repositories, and the processes through which Federal libraries develop digital collections.

Laura Drake Davis, a Senior State Records Archivist at The Library of Virginia, currently works with both electronic and print collections. Her interest is in the management of electronic records, particularly in strategies and techniques for processing and providing access to large quantities of electronic records and with special consideration for addressing the concerns associated with restricted and privacy-protected archival records.

1. INTRODUCTION: Bruce Fulton and Peter Botticelli (15 min.)
Beginning in 2007, SIRLS has been educating digital curators through its DigIn certificate program, which began through grant funding from IMLS. From the outset, the DigIn curriculum has
been built around hands-on, experiential technology learning combined with the practical and theoretical knowledge needed to manage a wide variety of digital collections in a fast-changing environment.

This agenda received a major boost in 2008, when SIRLS received grant funding from IMLS to design and test new methods for educating digital curators, and is led by Bruce Fulton, Jana Bradley, Peter Botticelli. For this project we are building a virtual laboratory environment and a range of exercises designed for use in online courses at the masters and graduate certificate levels. These exercises will be designed to help students learn the full set of functions needed to curate digital collections, as defined by the Digital Curation Centre through its Lifecycle Model [1].

In building the virtual lab, our intention is to design exercises that will help students learn the collection management process through realistic scenarios, using actual collection objects and current tools in use by practitioners. By gathering research data hope to demonstrate how experiential learning techniques can enhance students’ ability to master new technologies and better prepare them for careers in digital curation in a wide range of institutional and professional settings.

Thus, Bruce Fulton and Peter Botticelli of the University of Arizona will open the panel discussion by briefly describing our current agenda for research and curriculum development, including the initial set of virtual lab exercises we have planned for masters-level courses at the University of Arizona. We will also introduce the three key issues we plan to explore in the panel discussion:

2. Career Paths for Digital Curators (20 min.)

In designing a curriculum for digital curation, a critical issue is how to define the potential roles a curator might play in the field. As educators, we are actively seeking to match the career goals of individual students with the particular technology skills needed by digital curators. Thus, we will consider this issue both from the perspective of professional students interested in working as digital curators, and practitioners who currently employ digital curators.

3. Case examples of technology learning needs (20 min.)

After exploring technology learning issues from the individual point of view, we next consider institutional perspectives as we seek to understand how educational programs can help practitioners build the organizational capabilities needed for new types of data management and curation services. An important goal for the DigIn program is to identify up-to-date case examples to inform the curriculum. Thus, the panel will discuss curation activities and services in libraries (at the state, federal, and university levels especially). We also welcome input from other institutional perspectives, as we seek to define a set of case examples that show how the field is evolving at present.

4. The relevance of hands-on learning in digital curation (20 min.)

Here we will focus on some of the specific learning experiences we have developed for the DigIn program, and we will consider how they match the individual and organizational needs identified earlier. We will also consider ideas for new exercises using systems that are currently relevant for curation across the data management lifecycle. This supports our emphasis on using the most current and advanced technologies available for authentic, hands-on learning.

5. General comments and discussion (15 min.)

6. ACKNOWLEDGMENTS

Our thanks especially to IMLS, the University of Arizona, the Arizona State Library, Archives and Public Records, and the University of Arizona Libraries for their support of our research and teaching activities.

7. REFERENCES

Invited Demo Session II: Digital Curation Tools and Demos

David Pearson
Digital Preservation
National Library of Australia
Parkes Place, ACT 2600 Australia
dapearso@nla.gov.au

Chirag Shah
School of Info. & Library Science
Univ. of North Carolina at Chapel Hill
Chapel Hill, NC 27599 USA
chirag@unc.edu

Reagan Moore
School of Info. & Library Science
Univ. of North Carolina at Chapel Hill
Chapel Hill, NC 27599 USA
rwmoore@renci.org

Geri Bunker Ingram
OCLC Digital Collection Services
Seattle, WA, USA
ingramg@oclc.org

Andrew McHugh
University of Glasgow
Humanities Advanced Technology & Information Institute
Glasgow, Scotland
A.McHugh@hatii.arts.gla.ac.uk

Hans Hofman
Nationaal Archief Netherlands
2595 BE, The Haag,
The Netherlands
hans_hofman@planet.nl

ABSTRACT
This invited demonstration panel session brings together an international collection of tools for performing and facilitating digital curation in the practice setting, as well as for use in the education setting. This is the second of a two-part tools and demo session. Tools to be demonstrated in this session include: Prometheus, a digital preservation workbench [1]; Mediapedia, a prototype web-based resource on carriers [2]; ContextMiner, a framework to collect, analyze, and present contextual information along with the data [3]; iRODS (Integrated Rule-Oriented Data System), a data grid software system developed by the Data Intensive Cyber Environments (DICE) group and collaborators [4]; CONTENTdm, a complete software solution for the storage, management, and delivery of multi-format digital collections to the Web [5]; and DRAMBORA (Digital Repository Audit Method Based on Risk Assessment), a toolkit to facilitate internal audit by providing repository administrators with a means to assess capabilities, weaknesses, and strengths [6]. Additionally, four invited papers included in these Proceedings provide further information on Prometheus, Mediapedia, ContextMiner and CONTENTdm.

Keywords
Digital libraries, digital curation, data curation, digital preservation, planning, audit, evaluation, data grids, data management polices, digital repositories.

1. REFERENCES

This work is licensed under the Creative Commons Attribution-Noncommercial-No Derivative Works 3.0 Unported license. You are free to share this work (copy, distribute and transmit) under the following conditions: attribution, non-commercial, and no derivative works. To view a copy of this license, visit http://creativecommons.org/licenses/by-nc-nd/3.0/.

DigCCurr2009, April 1-3, 2009, Chapel Hill, NC, USA
Invited Demo: Prometheus: Managing the Ingest of Media Carriers

Nicholas del Pozo
Digital Preservation
National Library of Australia
Parkes Place, ACT 2600 Australia
ndelpozo@nla.gov.au

Douglas Elford
Digital Preservation
National Library of Australia
Parkes Place, ACT 2600 Australia
delford@nla.gov.au

David Pearson
Digital Preservation
National Library of Australia
Parkes Place, ACT 2600 Australia
dapearso@nla.gov.au

ABSTRACT
The National Library of Australia has a relatively small but important collection of digital material stored on common carriers such as floppy disks, CDs and DVDs. This includes both published material and unpublished manuscripts in digital form. In the past, preservation of the Library’s physical format digital collection has been taken care of manually, on a case-by-case basis, but this approach is insufficient to deal effectively with the increasing volume of material requiring preservation.

The Library has produced an application called Prometheus, which provides a semi-automated, scalable process for transferring data from carriers to preservation-managed digital storage. This is helping the Library to mitigate the major risks associated with storing the content on physical carriers: deterioration of the media and obsolescence of the hardware required to access them. Prometheus makes it easier to process the majority of carriers commonly encountered in the Library and to collect and manage metadata about their content. Although not perfect, Prometheus is helping the Library to save digital content before it is too late.

Keywords
Digital preservation, media carriers, National Library of Australia, obsolescence, open source software, Prometheus.

1. INTRODUCTION
The National Library of Australia has a relatively small but important collection of digital material stored on common carriers such as floppy disks, CDs and DVDs. This includes both published material and unpublished manuscripts in digital form. In the past, preservation of the Library’s physical format digital collection has been taken care of manually, on a case-by-case basis, but this approach is insufficient to deal effectively with the increasing volume of material requiring preservation.

The Library collects digital material through multiple acquisition streams and generally has little control over the physical format in which the material arrives. So, while most items fall into a small number of widely used carrier types, any long-term solution has to make provision for almost any kind of carrier, including carrier types which may not have been encountered yet. Moreover, this is a constantly growing problem; if we don’t deal with the digital materials that we have already collected, and ideally process new materials as a part of the acquisition process, accessing these carriers will soon become unmanageable, and eventually impossible.

Factors such as obsolescence and carrier degradation already make it difficult for digital preservation solutions to preserve access to digital content. Additionally, due to the potential volume and diversity of carriers and file formats, unless solutions are robust and semi-automated, the digital data that it is currently possible to preserve may not be. To avoid exacerbating the problem, it is key that solutions deal with current common carrier types as efficiently as possible, while providing access to, or a mechanism for preserving, as many older carriers as is practical.

2. PROMETHEUS
To ensure access to digital content on the most common carriers within the Library, the Digital Preservation Workflow Project produced an application called Prometheus. This application provides a semi-automated, scalable process for transferring data from carriers to preservation-managed digital storage. This is helping the Library to mitigate the major risks associated with storing the content on physical carriers: deterioration of the media and obsolescence of the technology required to access them. Prometheus makes it easier to process the majority of carriers commonly encountered in the Library and to collect and manage metadata about their content. Furthermore, Prometheus allows Library staff to link to catalogue records, create a byte-level image of the digital content, and transfer it to preservation-managed digital storage. Once the content is copied from the carrier, the integrity of the image is verified, and as much metadata as possible is harvested. Attaching a customisable ‘mini-jukebox’ (Figure 1) to a staff member’s workstation allows the accurate duplication of the content from a wider range of carrier types, such as USB thumb drives, memory cards or 3½ inch floppy disks. It also provides more reliable hardware for imaging CDs, and DVDs. The digital preservation section can use Prometheus to deal with carrier types that fall outside this range, such as 5¼ inch floppy disks, SyQuest disks or hard drives.

This work is licensed under the Creative Commons Attribution-Noncommercial-No Derivative Works 3.0 Unported license. You are free to share this work (copy, distribute and transmit) under the following conditions: attribution, non-commercial, and no derivative works. To view a copy of this license, visit http://creativecommons.org/licenses/by-nc-nd/3.0/

DigCCurr2009, April 1-3, 2009, Chapel Hill, NC, USA
The system incorporates a range of open source tools to undertake processing, including carrier imaging (dd [1], cdrdao [2]); integrity calculation and checking (Jaxsum [3]); file identification (DROID [4]); and metadata extraction (JHOVE [5], NLNZ Metadata Extraction Tool [6]). These tools are deployed using Java-based web services. Moreover, Prometheus has been designed in a modular way, so that tools and services can be easily upgraded or replaced as new versions are released or better software becomes available (Figure 2).

3. THE SOFTWARE RELEASED
Prometheus was designed for the Library’s specific environment, and therefore is not an ‘out of the box’ solution. However, it may be possible for other parties to use all or some of the requirements, other documentation or components. As such, the software has been released under the GNU General Public License V3.0. The latest version of Prometheus and its documentation is available from the project website [7]. A paper was presented on this project at the IFLA World Library and Information Congress in Quebec City, Canada, in August 2008 [8].

If we wait for the prefect system to be built, for the content on many carriers it will already be too late. Experience to date suggests that even though we all share the same fundamental problem, the sheer volume and diversity of carriers, as well as varying individual collecting and business environments, makes it unlikely that there will ever be a single software solution that can be used by everyone. At least for the Library, Prometheus provides a starting point to manage the ingest of, and preserve content from problematic and sometimes idiosyncratic carriers for long-term preservation, hopefully in a way that can advantage others.

This paper is based on the earlier paper, that appeared in Gateways Dec 2008 [9].

4. ACKNOWLEDGMENTS
Our thanks to Gerard Clifton, Snezana Mihajlovic and Joseph Mok, who worked with us on version 1.0 of Prometheus, and who continue with the development work for version 1.4.

5. REFERENCES


Invited Demo: Mediapedia: Managing the Identification of Media Carriers

Nicholas del Pozo
Digital Preservation
National Library of Australia
Parkes Place, ACT 2600
Australia
ndelpozo@nla.gov.au

Douglas Elford
Digital Preservation
National Library of Australia
Parkes Place, ACT 2600
Australia
delford@nla.gov.au

David Pearson
Digital Preservation
National Library of Australia
Parkes Place, ACT 2600
Australia
dapearso@nla.gov.au

ABSTRACT
All digital information is stored on physical carriers. Given the variations in carrier types, the quantity produced and in circulation, along with the potential importance of the content being stored on them, not taking any steps to document and preserve the characteristics of different carrier types will make it much more difficult, and eventually impossible, to extract content even in the short-term.

The Mediapedia is intended to provide a sustainable way of facilitating carrier type identification as well as documenting their technical requirements and general preservation information. By enabling a community of specialist individuals and organizations to collaborate in the documentation of these carriers it will hopefully create a sustainable body of knowledge which can be centrally and persistently accessed via the web. From a preservation and risk management perspective, we can either approach this problem as a community or ignore it at our individual peril.

Keywords
Digital preservation, media carriers, National Library of Australia, obsolescence, open source software, Prometheus.

1. INTRODUCTION
Anyone who has material stored upon obscure and older proprietary media carriers, and even more common carriers such as audio and video materials or floppy disks and CDs, will eventually encounter problems accessing this content. Although accessing current common carrier types may appear to be self evident presently, this may not always be the case. Over time, if organizations are not proactive there is a risk of losing access to the content stored on these carriers. In some cases, by the time an organization realizes there is a problem, it may already be too late to retrieve this content.

Even if there is an acceptance that access to a piece of media is ‘at risk’ of being lost, anyone trying to access its content will still be faced with the following issues:

- the carrier type may need to be identified;
- assuming the carrier type is known, then the technology (and associated dependencies for accessing those technologies) must be ascertained; and
- even when the above issues are resolved, accessing the media may still be problematic: technology (or parts of it) may not be readily available, or the carrier itself may have degraded so that it is no longer readable.

For example, 5¼ inch floppy disks are now a problematic carrier type to access (accurately or otherwise), though this was not always assumed to be the case. This is due to a number of factors such as short term deterioration of the physical materials, possible corruption of the data content on them and the loss in the availability of hardware (e.g., drives, cables and motherboard) and software (e.g., drivers and operating systems) which are required to load, recognize and read the physical disk. These factors are not necessarily mutually exclusive. Hardware, software, and file format obsolescence can occur independently from each other. In addition, not only can hardware become obsolete, but it can also be susceptible to chemical or physical degradation. For example, magnetic tape might start to de-laminate after a certain amount of time or after a certain amount of usage. Moreover, these problems are not only applicable to obsolete or older materials. Brand new carriers which are not yet in common usage may be just as inaccessible as older carriers that are no longer in use (e.g., HD-DVD).

Therefore, all carriers should be considered a temporary storage medium only. Ironically, in many cases these carriers have been perceived or marketed as long-term storage options. However, both the life-cycle of the carrier and the knowledge about it are dynamic. In the case of carrier specifications and documentation, their often ephemeral and proprietary nature means that while initially information may be readily available, it can easily disappear within a short period of time due to changing markets or business conditions. Information about older materials that pre-date the web is usually even more difficult to locate.

Because the problem is so diverse and complex, and the nature of carriers so dynamic, there is no single or simple solution. Therefore, there are implications to the types of carriers that content is stored on (both in the short- and long-term) that may not be immediately evident.
2. MEDIAPEDIA

In order to assist in managing these risks to carriers, we need to know a range of information about them. For example: when and by whom the carrier was created; when it was used; the advertised shelf-life versus the actual shelf-life; the requirements to access a specific carrier type. Mediapedia was designed to be an open, trusted and sustainable mechanism for documenting, retaining and disseminating this kind of knowledge [1]. The prototype of this web-based resource is intended to enable the identification of various types of carriers and their associated dependencies. Basic information which allows the identification of carrier types is provided, along with more detailed technical information about the carrier itself, and mechanisms that are needed to provide ongoing access. Future versions could include information about storage requirements, community based risk assessments and information about potential migration paths.

Unlike the Wikipedia [2], the Mediapedia does not require prior knowledge of a carrier’s name for discovery. It was specifically designed so that carrier types can be identified in a number of different ways. A user can search across other physical characteristics or descriptive details such as manufacturer, product code and other specific identifying markings that can be found on the carrier. For more advanced users, carriers can be identified through the use of a detailed and systematic classification system that was developed from several common standards, including Dublin Core Type Vocabulary [3] and the RDA/ONIX Framework for Resource Categorization [4]. The primary function of this classification system is to organize the carriers into meaningful and flexible taxonomical groupings or categories, and to make them discoverable to different audiences (see Figure 1). As such, the user can also search by Carrier or Process types within different Genres [5].

The Mediapedia doesn’t just store descriptive information about carriers, but also contains information about their dependencies and genre specific technical knowledge, and is designed to be both a human and machine harvestable resource. The data is intended to be curated and sustained by a base of trusted sources across a range of media genres. High quality images of each carrier type can be used to quickly confirm or refine search results (Figure 2). They can also be used as the basis for conducting a visual survey. This combination of a detailed classification system and the ability to search across multiple identification characteristics, attributes, descriptive text or images allows human users to quickly identify carriers. Machine users can harvest carrier information via persistent identifiers associated to each carrier type.

3. CONCLUSION

Given the variations in carrier types, the number of units produced and currently in circulation, and given the potential importance of the information being stored on them, not taking any steps to document and preserve this content is not an acceptable option. By identifying and knowing their characteristics and dependencies, we can more proactively manage the risk for these carriers, and therefore of the content that they contain.

It is hoped that the creation and use of the Mediapedia provides a sustainable way of facilitating carrier type identification as well as documenting technical and preservation information. Enabling a community of specialist individuals and organizations to collaborate in the documentation of these carriers will hopefully create a sustainable body of knowledge which can be centrally and persistently located. In addition, as this type of web-based service is not only human readable, but eventually also machine harvestable, it could potentially be re-used by other services and systems. From a preservation and risk management perspective, we can approach this problem as a community, or ignore it at our individual peril.

4. REFERENCES


Invited Demo: ContextMiner: Collect Different

Chirag Shah
School of Information & Library Science
University of North Carolina at Chapel Hill
Chapel Hill, NC 27599 USA
chirag@unc.edu

ABSTRACT
We present ContextMiner, a web-based service for collecting contextual information for digital objects from a variety of sources. ContextMiner lets one run campaigns that can include a set of queries that ContextMiner can run on various sources, such as YouTube and blogs, and keep extracting and adding contextual information to the collected objects based on their usage. Such contextual information can help to make sense of digital objects and better preserve them.

Categories and Subject Descriptors
H.3.7 [Information Storage and Retrieval]: Digital Libraries—Collection; H.5.3 [Information Interfaces and Presentation]: Group and Organization Interfaces—Web-based interaction

Keywords
Digital curation, digital preservation, contextual information.

1. INTRODUCTION
ContextMiner is a framework to collect, analyze, and present contextual information along with the data. It is based on an idea that while describing or archiving an object, contextual information helps to make sense of that object or to preserve it better [1, 2]. This idea has been realized as a web-based service, called ContextMiner,1 that provides tools to collect data, metadata, and contextual information off the web by automated crawls (Figure 1).

ContextMiner helps one (1) run automated crawls on various sources on the web and collect data as well as contextual information, (2) analyze and add value to collected data and context, and (3) monitor digital objects of interest over a period of time.

2. USING CONTEXTMINER
Once a user signs up for a free account, he/she can immediately start creating campaigns. A campaign in ContextMiner is a project that is based on running several automated processes and collecting data, metadata, and contextual information. Following is a typical flow of using ContextMiner:

1. Start a new campaign based on some story, concept, or an object.
2. Choose the sources (Web, Blogs, YouTube) that you want ContextMiner to do your searches and crawls on.
3. Once you provide all the required parameters, ContextMiner can immediately start running your campaign. You can access all your campaigns and collected data as well as contextual information through its website (Figure 2).
4. You can manipulate individual items as well as related items that are collected by the above processes to add your interpretation and meaning to the campaign.

Figure 1: ContextMiner architecture

---

1 http://www.contextminer.org
3. USAGE EXAMPLE

Let us now look at an example of capturing contextual information with ContextMiner. One of the sources ContextMiner works with is YouTube. While YouTube provides many valuable attributes relating to a video, we may need to explore other sources such as blogs to complete the picture [3]. For instance, look at one of the most popular (viral) videos on YouTube: ‘Vote Different’. To many people it is not clear where it came from--what the story is behind, who created it, and why. A screenshot of this item collected from YouTube by our system is shown in Figure 3. Some of the basic information about this video, including description, author name, and keywords, can also be seen.

One of the kinds of contextual information that ContextMiner captures is the links from other webpages on the web (inlinks) to a given digital object. Now if we look at the in-links collected to this YouTube video (Figure 4), we see that one of the articles linking to the above video talks about the author of this video. As we look at this article webpage, we can see that it talks about who created this video, why, and what is the background for the video. We can also find the original ‘Think Different’ video embedded in the article. Together, these objects provide us well enough contextual information to document the given digital object in a more meaningful way.

4. CONCLUSION

We have been using the ContextMiner framework and services for harvesting videos and contextual information relating to the presidential elections 2008 [4]. In addition to this, we have also been running crawls for collecting data and contextual information on a variety of topics, such as energy, epidemics, health, natural disasters, and truth commissions.

At the time of writing this, there are more than 200 users who have been using ContextMiner for several months, and have collected millions of objects (YouTube videos, blogs) and related contextual information. ContextMiner is also in use by several members of the National Digital Information Infrastructure Preservation Program (NDIIPP) and can be used by teachers or others who wish to harvest content on specific topics. Further development providing access to more sources, and tools for information exploration is underway. ContextMiner is available as open source code or a web-based service from http://www.contextminer.org.

5. REFERENCES


Invited Demo: Teaching with CONTENTdm in the Digital Curation Curriculum

Geri Bunker Ingram  
OCLC Digital Collection Services  
Seattle, WA 98119 USA  
ingramg@oclc.org

ABSTRACT
One challenge for LIS programs is to develop digital curators who demonstrate a sound grasp of best practices and core principles in information management as they make use of modern digital collection management tools. Selection and description, presentation and preservation—at every stage in the lifecycle of electronic materials management, no less than with traditional librarianship, highly refined curation skills begin in the LIS program. Except for those curricular tracks that focus on tool design and development, library and information studies programs must leverage as-built curation tools as they teach core concepts—acquisition, cataloging, information retrieval, and preservation of the record of knowledge. During this session, we will look at how faculty are meeting their teaching and learning objectives while leveraging CONTENTdm in the LIS program.

Categories and Subject Descriptors
H.2 [Database management]: H.2.4 [Systems] distributed databases, multimedia database; H.2.7 [Database administration]; H.2.8. [Database applications] image database; K.3.2 [Computer and Information Science Education] curriculum, information systems education, literacy

Keywords
Digital asset management; electronic resource management; digital curation; curriculum

1. INTRODUCTION
Today’s librarians are managing a universe of data that is growing at an unfathomable rate. To master the art and science of digital curation one must learn quickly how critical data become information, and how information is synthesized into knowledge.

Daily we grapple with the responsibility of preserving this exploding record of knowledge for posterity. What is the size of the digital universe that libraries and archives must manage, and how fast is it growing? In 2007, for the first time, the amount of electronic information created exceeded the available storage. In 2007 – the size of the digital universe was 281 EXABYTES (281 billion GIGABYTES) By 2011 the universe will be ten times larger than it was in 2007. [1]

OCLC
OCLC, a not-for-profit global cooperative, is on a mission to connect people to knowledge through library cooperation. It provides a sustainable infrastructure for WorldCat, the world’s largest digital repository, and makes it publicly and freely available at http://worldcat.org. OCLC members work together on local, regional and global bases, using collaborative methods and interoperable tools to select and describe, present and preserve both metadata and actual digital objects in this digital repository. Whether through sharing physical and digital resources or though cooperative cataloging, librarians have used OCLC tools for decades to deliver top quality services to researchers.

OCLC Digital Collection Services
Collaboration is at the heart of the enterprise, and sharing OCLC tools amplifies the effort and skills of catalogers and curators who provide top quality resources to researchers around the world. The mission of OCLC Digital Collection Services (DCS) is to offer services and products is to

- Build a global digital repository within the OCLC cooperative;
- Enable the creation of institutional digital repositories.
- Provide long-term archiving of high resolution materials.
- Work with digital library leaders to develop and evolve new and best practices for digital collections.

Practical applications in the curriculum
OCLC has for decades provided the opportunity for MLIS candidates to use industry-standard tools in the classroom—whether for cataloging, resource sharing, or reference work. OCLC DCS is the home of CONTENTdm, Connexion digital import, the Web Harvester, the Digital Archive, and WorldCat Metadata Harvesting. Since 2004 we have freely provided CONTENTdm to LIS programs worldwide; in addition, we host the LIS teaching and student collections at OCLC’s data center in Dublin, Ohio.

CONTENTdm is described in: Digital Library Curriculum Development 1 Module: 5-b: Application Software 2
CONTENTdm Digital Collection Management software was conceived (as “CONTENT”) in the 1990’s by the Center for Information Systems Optimization (CISO) at the University of Washington. The original developers at CISO, under the leadership of Dr. Greg Zick, formed a new company in 2001—DiMeMa. OCLC acquired the company in 2006, and the original developers continue to improve CONTENTdm today as OCLC Digital Collection Services.

CONTENTdm is used to manage millions of digital items in libraries and cultural institutions around the world. Collections can be accessed through standard web browsers, and is highly interoperable with local systems. Metadata can be exported to OCLC WorldCat, and CONTENTdm servers function as OAI data repositories harvested around the world. [2]

CONTENTdm 5, to be released in Fall Quarter 2008, will substantially advance the building of the largest digital repository in the world. WorldCat’s Find search engine has been integrated into CONTENTdm, bringing faceted searching and full support for Unicode. It will feature streamlined collection curation workflows and powerful reporting for assessment and cost-effective electronic resources management. Sophisticated authority control tools will feature the incorporation of nine widely adopted thesauri—

- Art & Architecture Thesaurus (AAT)
- Dublin Core Metadata Initiative Type Vocabulary
- Getty Thesaurus of Geographic Names (TGN)
- Guidelines On Subject Access To Individual Works Of Fiction, Drama, Etc., 2nd ed., form and genre
- Māori Subject Headings / Ngā Ūpoko Tukutuku
- Medical Subject Headings (MeSH®) 2008
- Newspaper Genre List
- Thesaurus for Graphic Materials
- Union List of Artist Names (ULAN)

Case studies
Drawing on the experience of dozens of LIS programs using CONTENTdm in the classroom, we will discuss:

- In what specific ways does teaching with CONTENTdm meet the needs of the digital curation curriculum? E.g., providing an easy to implement system for learning metadata standards, crosswalking and creating application profiles
- How does it serve graduate education and continuing education for practitioners?
- How has teaching with CONTENTdm in the LIS program helped faculty to prepare students with emerging professional competencies and personal attributes for employment in digital curation environments?

2. REFERENCES

Working with powerful, industry-standard, and practical tools in the information management curriculum prepares skilled curators, ready to meet professional challenges inherent in this exploding digital universe.
ABSTRACT

This panel will present a variety of continuing education initiatives in digital curation, digital preservation, and data curation. Panelists will discuss the potential intersections of these initiatives and the opportunities for enhancing synergies and extending resources. The speakers represent curriculum development from a range of digital disciplines and organizational settings and will discuss the implications for sustainability of educational programs within these settings.

Categories and Subject Descriptors
H.3.4 [Information Storage and Retrieval]: Systems and Software; H.3.7 [Information Storage and Retrieval]: Digital Libraries

Keywords
Digital curation, digital preservation data curation, training, education.

1. INTRODUCTION

The panel will address four core topics: the digital disciplines in which curriculum for training and education are being developed; the educational objectives and scope of current curriculum development initiatives; the intersections between these initiatives and the opportunities for collaboration; and the implications for sustainability of educational programs within a range of settings.

2. DIGITAL DISCIPLINES

First, the panel will consider the emergence of a range of digital disciplines: digital preservation, the lineage of which can be traced to the 1996 *Preserving Digital Information* report; data curation, an established field developed over decades of practice in the sciences that pertains to the value-added processes and services to make data as meaningful and understandable as possible; and digital curation, a new field that emerged within the past decade that focuses on managing digital assets across their lifecycles.

3. CURRENT INITIATIVES

This section reviews the scope, educational objectives, and intended outcomes of a variety of continuing education initiatives in digital curation, digital preservation, and data curation.

3.1 DigCCurr Profession Institutes

With funding from the Institute for Museum and Library Services (IMLS) the School of Information and Library Science (SILS) at the University of North Carolina at Chapel Hill (UNC-CH) is partnering with the National Archives and Records Administration US (NARA) in DigCCurr II: Extending an International Digital Curation Curriculum to Doctoral Students and Practitioners. Helen Tibbo will discuss the DigCCurr II project that seeks to develop an international, doctoral-level curriculum and educational network in the management and preservation of digital materials across their life cycle. This project will also provide professional institutes for cultural heritage information professionals already working in this arena. The first Professional Institute, "Curation Practices for the Digital Object Lifecycle" will be held June 21-26, 2009 with a follow-up session January 6-7, 2010 in Chapel Hill, NC. Each day of the June session will include lectures, discussion and a hands-on "lab" component. This year's international teaching staff includes Christopher Lee, Richard Marciano, Helen Tibbo, and Carolyn Hank from the University of North Carolina at Chapel Hill; Seamus Ross of the University of Toronto; Manfred Thaler, University of Cologne; and Nancy McGovern, University of Michigan.

3.2 Data Curation Education Program (DCEP) Summer Institutes

Data curation is the active and on-going management of data through its lifecycle of interest and usefulness to scholarship, science, and education. Data curation activities enable data discovery and retrieval, maintain its quality, add value, and
provide for re-use over time, and this new field includes authentication, archiving, management, preservation, retrieval, and representation. Melissa Cragin will discuss the DCEP Summer Institutes at the Graduate School of Library and Information Science (GSLIS) at the University of Illinois at Urbana-Champaign (UIUC). DCEP will hold a Summer Institute for Humanities Data Curation from May 18 through 22, 2009. Allen H. Renear, Associate Dean for Research at GSLIS, will direct the Institute and sessions will be conducted by leading specialists in the various fields that are involved in the curation of humanities data.

3.3 Digital Preservation Management Workshop and Tutorial

The Inter-university Consortium for Political and Social Research (ICPSR) is pleased to be the host institution for the Digital Preservation Management Workshop and Tutorial. Nancy McGovern will discuss this expanded program, which is based on the workshop curriculum initially developed at Cornell University and supported with funding from the National Endowment for the Humanities (NEH). The workshop series is intended for managers who are or will be responsible for digital preservation programs in libraries, archives, and other cultural institutions. The goals of the workshop are to foster critical thinking in a technological realm and provide the means for exercising practical and responsible stewardship of digital assets in an age of technological uncertainty. The next workshop is scheduled for May 3-8, 2009. Future 5-day workshops are scheduled for October 11-16, 2009 and May 2-7, 2010.

3.4 Digital Curation 101

With funding from the Joint Information Systems Committee (JISC), the Digital Curation Centre UK has developed the Digital Curation 101 workshop series to assist researchers in developing sound data management and curation plans. Joy Davidson will discuss this new series from which workshop participants will gain an understanding of the range and nature of data management and curation activities that should be considered when planning new research projects, and will be better equipped to develop and implement sound data management and curation plans for future research funding bids. The target audience for these workshops is researchers with funding body data management and curation mandates to fulfill and information management specialists. A key goal is the integration of these communities of practice to share their experiences and to identify where, when and how they could best cooperate to meet data curation challenges.

3.5 Digital Preservation Europe (DPE)

With funding from the European Commission, DigitalPreservationEurope (DPE) fosters collaboration and synergies between many existing national initiatives across the European Research Area. DPE addresses the need to improve coordination, cooperation and consistency in current activities to secure effective preservation of digital materials. DPE, building on the earlier successful work of ERPANET, facilitates pooling of the complementary expertise that exists across the academic research, cultural, public administration and industry sectors in Europe. Hans Hofman will discuss the training courses that are organized by DPE in close collaboration with other European projects, such as Planets, Caspar, nestor, Shaman, under the banner of WePreserve.

4. INTERSECTIONS

This section discusses the potential for combining, sharing, extending, adapting, and formalizing next steps for curriculum developments in these areas. It seeks to maximize synergies and a wide variety of perspectives and expertise.

5. SUSTAINABILITY

This section probes various paths to sustainability for these educational offerings and the role that national funding agencies, national libraries and archives, and graduate educational programs might play in this effort.

6. CONCLUSIONS

Preparing practitioners, managers, faculty, students, and researchers to continually address the challenges of ensuring the longevity of digital content is a responsibility of all of these disciplines and initiatives. It would benefit the communities that are concerned with digital longevity issues if these curriculum development initiatives collaborated to define common goals and objectives, to enable shared and modular curriculum, and to encourage cooperation.

7. ACKNOWLEDGMENTS

The curriculum development initiatives discussed in this panel were funded by the European Commission, IMLS, JISC, and NEH.

8. REFERENCES

Snapshot of Digital Preservation in Federal Libraries

Deborah Keller  
U.S. Army Corps of Engineers, Humphreys Engineer Center Support Activity Library  
Alexandria, VA 22315 USA  
deborah.eb.keller@us.army.mil

Patricia Murphy  
U.S. Department of Agriculture, National Agricultural Library  
Beltsville, MD 20705 USA  
pmurphy@nal.usda.gov

Christina Dunn  
U.S. Department of Education, National Library of Education  
Washington, DC 20202 USA  
christina.dunn@ed.gov

Claretta Crawford  
U.S. Army, Fort Leonard Wood  
Bruce C. Clarke Library  
Fort Leonard Wood, MO 65473 USA  
claretta.crawford@us.army.mil

ABSTRACT
In this panel, we represent the variety within the federal library sector as seen through the digital projects being pursued. Panelists include librarians and digital project coordinators from both large and small federal libraries containing diverse subject collections, and address the challenges presented by digitizing different physical media. Each panelist’s comments will focus on three areas: First, the presenter will frame their comments by providing the purpose or impetus for pursuing digital preservation project in their library. Second, the presenter will discuss approach the library has taken in pursuing the project, including considerations such as scanning, repository software, and other technical requirements, as well as the availability and uses of human and financial resources. Finally, the presenter will analyze the project, highlighting steps or decisions that have led to successes and pointing out areas of difficulty that have been overcome as well as those that present ongoing challenges in the project. A moderator will facilitate discussion between panelists and field questions from the audience. Although federal libraries support organizations whose missions and areas of interest differ considerably, this snapshot will illustrate that their digital projects experience many of the same constraints and that regardless of the library or project size, librarians often are faced with similar challenges.

Categories and Subject Descriptors  

Keywords  
Federal libraries, digitization, preservation, project management, resource management, knowledge management.

This work is licensed under the Creative Commons Attribution-Noncommercial-No Derivative Works 3.0 Unported license. You are free to share this work (copy, distribute and transmit) under the following conditions: attribution, non-commercial, and no derivative works. To view a copy of this license, visit http://creativecommons.org/licenses/by-nc-nd/3.0/.

DigCCurr2009, April 1-3, 2009, Chapel Hill, NC, USA

1. INTRODUCTION
The Federal Library and Information Center Committee (FLICC) Preservation and Digitization Working Group is an ad hoc working group of library professionals that has been organized to develop strategies for the long-term preservation and access to federal library resources. On behalf of that working group, this panel intends to present a snapshot of digital preservation efforts underway within federal libraries today.

The federal library community is a diverse one. Libraries provide support to agencies with missions as disparate as scientific research is from law and justice. Like their varied missions, libraries also receive differing support from their parent organizations. National libraries with missions to support federal departments may be large and well resourced, like the Library of Congress, but spending cuts continue to threaten libraries in the Environmental Protection Agency and Department of Agriculture, among many others. Many smaller libraries continue to support agencies such as the U.S. Geological Survey, the General Accountability Office, or the Peace Corps with more narrowly defined missions.

2. METHOD
This panel presentation will feature librarians from four federal libraries currently involved in a digitization project as well as a fifth librarian who will serve as the discussion monitor. Each panelist will give a 10-12 minute presentation about their library’s digital project that will address three main issues. First, they will discuss the purposes of their digital projects, what they hope to accomplish, and how they intend to provide benefit to the community or communities that they serve. Second, the panelists will discuss at length the approaches that their libraries have used in digital project efforts. Their comments will address both the tangible and intangible organizational sensing that takes place when planning a project, how technical concerns have played into the path that the project has taken, and the solutions that they have found in particular processes, workflows, software, or services. Third, panelists reveal the points at which they celebrated milestones in their projects. They also discuss the obstacles, limitations, and challenges which impacted their project, how these have been overcome, or how the project team is working to change their influence. A moderator will highlight
similarities between libraries, facilitate discussion between panelists, and direct questions asked by interested audience participants.

3. DISCUSSION

The services and initiatives of federal libraries are influenced by two factors: the mission of the department or agency that they support and the information needs of the taxpayers. In most libraries, these factors are not equal concerns and may, in some cases, conflict with one another. The motivations behind pursuing a digital project, therefore, are often complex. In many libraries, digitization is pursued as a means of preserving valuable content, improving the access to research materials, and reducing the square footage occupied by physical materials.

The way that digital projects are approached by individual libraries is the product of many factors. Some of these, like the goals of a digital project, can often be clearly defined. Time, financial resources, and personnel available for a project can be quantified, if not always easily. However, many factors are more intangible. For example, what is the political climate of the organization, particularly its support for library initiatives? What kinds of computing or networking resources does the organization have and how responsive is the IT staff to the needs of its customers? How concerned is the organization with the security of its information? How important is it that the information be preserved for long-term use?

Curation of digital projects in federal libraries also takes into consideration a variety of issues. Digital libraries present ongoing collection development concerns. Some libraries focus on continual acquisition of new materials or expanding their digital projects to include increasingly larger percentages of digitized legacy materials. Others focus most on integrating these digital collections with legacy physical formats is often challenging, as is synchronizing metadata schemes and access methods. Traditional areas of library “shortage,”—funding, personnel, physical space, and technical support—require ongoing monitoring in most settings because of the regular changes that occur in these areas. Finally, outsourcing, of scanning services, metadata creation, software hosting and even library personnel, is especially common in digital projects and further complicates the management of these projects.

4. CONCLUSIONS

Regardless of their differences, focusing on one aspect of library services, management of digital projects, reveals that federal libraries share important characteristics that affect their daily operations. Nearly all federal libraries feel that they do not have sufficient resources to pursue their digital projects. For some, this is felt as a “cash insufficiency,” being unable to pursue digitization as rapidly as they believe are necessary for access or preservation purposes. For others, it is a personnel insufficiency, having too few staff or staff who must devote considerable time to other projects, therefore slowing down digital project efforts. Many feel that they have insufficient technical support. For some, digital projects require understanding technical details and programming languages that they do not already use. For others, the agency’s information technology staff is unable to provide the level of support that their digital projects require. Many libraries find that their organizations do not understand the costs involved in long-term digital projects and are unwilling to provide sufficient support to pursue them.

However, an increasing number of libraries believe that pursuing digitization is so important that they squeeze these projects into existing library programs and budgets. Library patrons increasingly expect that library resources will be accessible to them from their desktops and from home. Librarians recognize that preserving their materials, whether scientifically significant, historical, or simply the legacy of their organization, has growing importance to the library’s role in organizational knowledge management efforts. In a political climate where federal employees at all levels are “doing more with less,” federal libraries are taking steps now to manage their resources most efficiently and effectively for the long term. While the details often vary, common themes and concerns are easily found among federal libraries.

5. ACKNOWLEDGMENTS

Our thanks to the Federal Library and Information Center Committee (FLICC) for facilitating cooperation and information sharing between federal libraries and its support for outreach beyond the federal library community.
Digital Curation Vignettes: Personal, Academic, and Organizational Digital Information

Joan E. Beaudoin  
iSchool  
Drexel University  
Philadelphia, PA 19104 USA  
jeb56@drexel.edu

Maria Esteva  
Texas Advanced Computing Center  
University of Texas at Austin  
Austin, TX 78758 USA  
maria@tacc.utexas.edu

Andrea Japzon  
iSchool  
Drexel University  
Philadelphia, PA 19104, USA  
acj26@drexel.edu

ABSTRACT
This panel presents variations on the theme of digital curation by examining the digital information management and preservation practices of three different populations. Personal digital information management, personal collections transferred to institutional repositories, and a digital archiving case in a private organization, offer a wide view of the types of contexts in which digital material is being produced “in the wild.” Across the cases we found that digital record-keeping and preservation practices are not well understood or established, and that a vast amount of digital content created currently is at risk. Other issues, such as an individual’s perception of digital information value, and the feasibility of preservation beyond an individual’s or organization’s lifetime, surfaced as determinants of the current situation. The findings have important implications for appraisal and post-custodial archival strategies. They are also useful for identifying critical decision points when digital curation issues are best addressed.

Keywords
Digital curation, institutional repositories, post-custodialism, hybrid archives, digital images, digital preservation, information value

1. PERSONAL DIGITAL INFORMATION
Libraries and other memory institutions are aware of the crisis in digital preservation and are taking steps to preserve our collective cultural heritage [1, 2]. In contrast, the research of Marshall et al. [3] suggests that individual consumers are much less aware of the impermanent state of their digital possessions, or if they are aware they feel disempowered to do anything about it. As a result, valuable representations of personal memories intended for future generations will be lost through ignorance and/or benign neglect [4], and representations of family and social histories will be lost to what has been called the "digital dark ages" [5].

Findings from a qualitative research study on the behaviors that contribute to the preservation of personal digital information, involving 26 participants recruited from public library friends groups will be presented. The theory of information source horizons is used to explore the continuum of physical and digital information, and to elicit the criteria participants use to determine the value of their personal information [6, 7, 8].

Discussion will include what types of personal digital information individuals are saving, where they save it, and why. The value given by individuals to their personal digital information is explored in relation to information management behaviors, tangible versus intangible information, and storage location selections. This research will contribute to the understanding of affective and cognitive interactions with personal digital information; the values associated with it; and the gaps in the knowledge possessed by individuals regarding digital preservation.

2. DIGITAL IMAGE COLLECTIONS OF ACADEMICS
It is broadly understood that visual information is critical to the pedagogy and scholarship of many academic disciplines. Therefore it seems reasonable to expect to find coverage of the digital image collections amassed by faculty to support their teaching and research in the literature surrounding digital curation and preservation. While discussions of the archival and preservation practices surrounding digital images are present [9, 10, 11] the importance of these images as documents of our greater cultural heritage and an institution’s curriculum and its faculty’s scholarship remains unacknowledged. By and large institutional repositories have been concerned with scholarly and institutional publications, with collections of images and other teaching materials assuming a sometimes uneasy place within the repository. This is an unfortunate situation considering these collections will provide future research opportunities [12] as well as providing much needed faculty interest and support for traditionally under-utilized institutional repositories [13, 14, 15, 16, 17].

A qualitative study examines the preservation practices connected to the personal digital image collections of six academic art historians and six archaeologists at several academic institutions. Through semi-structured interviews and observations the participants’ attitudes toward their digital image collections will be presented alongside the preservation strategies they employed. The participants’ self-perceived technological skills and their awareness of institutional support for their efforts are also discussed. How faculty view the role of the institution in maintaining their image collections is explored through topics
such as their awareness of institutional repositories and the availability of staff to help them manage and preserve their images. The research provides an account of the preservation needs and current practices surrounding the digital image collections of art historians and archaeologists. The findings of this study have implications for the development of curatorial practices to cope with personal academic digital image collections maintained by individuals.

3. ARCHIVING THE ELECTRONIC RECORDS OF A PRIVATE ORGANIZATION

A four year experience archiving the records of a philanthropic organization in Argentina revealed the possibilities and limitations of post-custodial approaches, in which the archive creator is responsible for the long-term custody of his archive, and the archivist guides the archiving process [18]. The case study archive originated in mid 1980’s as a centralized paper record-keeping system, evolving over the years into a hybrid of networked databases, a semi-centralized paper file, and electronic records stored in a shared directory. Research of local legal and archival regulations highlighted a dearth of recommendations concerning what electronic records to retain in the context of hybrid information systems, and of digital preservation best practices to follow during the retention period that precedes the closure of private organizations. Access and preservation were further complicated by lack of systems administration documentation and electronic record-keeping consistency. All of this created doubts for the administration about the feasibility of preserving the digital information and pointed instead toward keeping only the paper files.

To address the problems presented above, the archiving process included designing a post-custodial digital preservation strategy and discerning the functions of the paper and the digital information systems. Studying the archive’s formation process [19] allowed recovering information about the technologies used over 20 years to create and manage electronic records. The study suggested a digital preservation strategy based on archival acquisition protocols, system’s administration best practices, migration on demand [20], and virtual migration [21].

The functions of the paper and electronic records were explored with text mining, animated visualization, social network analysis, and staff member’s ethnographies [22]. Results suggest that the electronic records can show work-processes and organizational dynamics, and that the paper records best reflect the institution’s actions and accomplishments. It was concluded that both information systems should be preserved.

4. REFERENCES


Cooperative Approaches to Digital Preservation: Panel

Martin Halbert
University Library
Emory University
Atlanta, GA 30322 USA
martin.halbert@emory.edu

Richard Pearce-Moses
Arizona State Library, Archives and Public Records
Phoenix, AZ 85007 USA
rpm@lib.az.us

Tyler Walters
Georgia Institute of Technology
Atlanta, GA 30332 USA
tyler@gatech.edu

Aaron Trehub
RBD Library
Auburn University
Auburn University, AL 36849 USA
trehuaj@auburn.edu

Jonathan Crabtree
Odum Institute for Research in Social Science
University of North Carolina at Chapel Hill
Chapel Hill, NC 27599 USA
jonathan_crabtree@unc.edu

ABSTRACT
In this panel, representatives from four archives – the MetaArchive Cooperative, Alabama Digital Preservation Network, Data-PASS, and the Persistent Digital Archives and Library System – discuss the versatility, low cost, and compelling benefits of using cooperative distributed digital preservation networks to safeguard categories of digital content that define our culture, identity, and history and that might otherwise be lost as a result of natural disaster, human error, or neglect.

Categories and Subject Descriptors
H.3.4 [Systems and Software]: Distributed Systems, Information Networks.
K.6.4 [System Management]: Centralization/Decentralization

General Terms
Management, Design, Economics, Reliability, Legal Aspects, Verification.

Keywords
Distributed Digital Preservation

1. INTRODUCTION
Participants in this panel will describe the forces that drove them to preserve their institutions’ unique collections using a distributed digital preservation approach called PLNs (Private LOCKSS Networks), and the ways they were able to adapt their networks to meet their institutions’ specific needs. The discussion will address the versatility, low cost, and benefits of using a distributed digital preservation network to safeguard categories of digital content that define our culture, identity, and history and that might otherwise be lost as a result of natural disaster, human error, or neglect.

Each panelist represents an institution deeply committed to safeguarding our cultural heritage for future generations. Researchers at each institution discovered the PLN to be a flexible, effective tool with which to accomplish their goals. Additionally, they discovered the inherent value of building networks with other like-minded institutions. Each panelist was able to encourage other libraries and archives to collaborate in building an infrastructure that would safely and reliably protect assets and records of national significance.

The panelists will describe how their networks differ from the others. Examples of these differences include:

- single-state vs. multi-state model
- different business strategies and pricing models
- technology and work flow integration decisions (including hardware needs and network management)
- types of materials being preserved (range from digital collections about the American South to electronic public records)

Distributed digital preservation networks built on LOCKSS’ open-source technology have been growing in popularity among research and government institutions that wish to collect and preserve materials of high importance to specific communities.

2. TECHNICAL ISSUES
While the technologies behind PLNs are affordable and standard, much work has taken place to date to optimize these networks’ performance. In some cases, new technologies have been created—as in the case of the MetaArchive Cooperative and its conspectus database, which is now available as an open source module to be used with the LOCKSS software. Some of these networks are pushing technological limits in regards to the size and number of collections (or “archival units”, in the LOCKSS vernacular) that are being harvested. Many other issues will be discussed by the panel, including issues relating to indexing network holdings, ingest challenges, explorations in format migration and emulation, networking challenges such as open port...
management, overseeing and ensuring the successful multiple node harvest of collections, and the overall aligning of technologies such as hardware, operating systems, and middleware to ensure that nodes are successful in their interactions and communications.

3. SUSTAINABILITY ISSUES
The cornerstone reason for founding the organizations behind PLNs is to sustain the preservation network and its capacity to maintain and manage its content over long periods of time. The experiences of the PLNs on this panel will be discussed with a view to how they have approached the need to sustain their organizations to date. Approaches range from making use of existing state consortia (ADPNet), to grafting distributed digital preservation management onto multi-institutional consortia (Data-PASS), to creating new multi-institutional multi-state consortia (PeDALS), to creating a brand new cooperative with its own independent management entity (MetaArchive Cooperative). Financial issues are key to sustainability, and each cooperative will expand upon its approach to garnering adequate finances to sustain preservation activity, including working to maintain extremely low costs to participation, managing the number of members, and devising an annual dues structure to support and sustain cooperative business.

4. ORGANIZATIONAL ISSUES
Each cooperative has faced specific challenges in initiating, growing, and maintaining its multi-institutional organization. Foundational issues have been addressed, such as establishing governance and related documents, defining rights and responsibilities of the participating institutions, financial arrangements, technology specifications, and processes for adding new as well as withdrawing members and collections. Panelists will share their cooperatives’ experiences and the best practices they have derived from their organization-building activities. The emerging theme from this work is the development of formalizing trust relationships in new digital preservation federations like those being represented in this panel session.

5. REPOSITORY ISSUES
PLNs are beginning to be explored with regard to preserving the content of popular, open source, institutional repository systems such as DSpace and EPrints. Consortia in the U.S., Italy, the U.K., and South Africa have expressed interest in or are currently exploring the establishment of PLNs to harvest and preserve their repository content. One of the largest DSpace repositories in the world is Georgia Tech’s SMARTech repository, which is being harvested as part of the MetaArchive Cooperative’s network. The MetaArchive – DSpace experience will be examined for early lessons learned. Panel members will discuss how the process is being refined, anticipate improvements that can be made with new protocols becoming available, and share best practices currently being documented for LOCKSS harvesting of repository systems.

6. ACKNOWLEDGMENTS
We would like to acknowledge the support of the Library of Congress National Digital Information Infrastructure and Preservation Program (NDIIPP), Institute of Museum and Library Services (IMLS), and the National Archives and Records Administration National Historical Publications and Records Commission (NHPRC).

7. REFERENCES
Extending the Data Curation Curriculum to Practicing LIS Professionals
Graduate School of Library and Information Science
University of Illinois at Urbana-Champaign
Champaign, IL, 61820 USA
{cragin, lcsmith, clpalmer, pheidorn @illinois.edu}

ABSTRACT
In this panel we will present an overview of and outcomes from the inaugural Summer Institute on Data Curation held at the Graduate School of Library and Information Science at the University of Illinois at Urbana-Champaign. The Institute addresses a growing need for continuing professional development in data curation. Panelists will present their experiences attending the Institute, and discuss these in relation to the current and ongoing data curation activities at their own universities.

Categories and Subject Descriptors
K.3.2 [Computing Milieux]: Computer and Information Science Education - Curriculum

Keywords
Data curation, library and information science professionals, university IT professionals, in-service training

1. Summer Institute on Data Curation
The growing need for academic and research libraries to participate, and in some arenas lead, the work of stewarding scientific research data is placing new demands on libraries. While this may seem like a distracting burden to libraries already straining to “retool” in response to changes in scholarly communication, it is very much in accord with these activities. Engaging in data curation work will provide opportunities for broad participation in the growth and support of e-Science and e-Research across the university campus, but a trained workforce will be needed for effective engagement with scientists and scholars (ARL, 2007). To address the need for skilled librarians, the Graduate School of Library and Information Science (GSLIS) at the University of Illinois at Urbana-Champaign, with support from the Institute of Museum and Library Services, is pioneering the development of data curation education for library and information science professionals. As a part of this program, and in response to growing demand from practicing Library and Information Science professionals, the GSLIS held in June, 2008, its inaugural Summer Institute on Data Curation. Twenty-nine LIS professionals and one LIS professor attended the Institute. This group represented a range of academic and research institutions, including 14 U.S. universities, the Woods Hole Oceanographic Institution, and the Canada Institute for Scientific and Technical Information.

Data curation is the active and on-going management of research data through its lifecycle of interest and usefulness to scholarship, science, and education. Data curation “goes beyond that of enhanced present-day re-use, and of archival responsibility, to embrace stewardship that adds value through the provision of context and linkage: placing emphasis on publishing data in ways that ease re-use and promoting accountability and integration” [2]. Here is how the GSLIS approach to data curation education and research is unique: At its core, data curation concerns the application of LIS and Archival theory and tools to maximize the current and future usefulness of research data. Above technical and computing tools, we give primacy to the value-added services that will be required for collection development, representation and linking, discovery, access and re-use. In addition, we bring a theoretical perspective that partners us with the scientists and scholars in their research process, to support data management and scholarly communication needs from the beginning of the research process, to best facilitate data organization, preservation, and re-use.

The over-arching goal of the Summer Institute on Data Curation was to broaden participants’ understanding of data curation (form, function, activity) in the academic library context. Objectives included increasing participants’ understanding of skill sets, resources, and collaborations that are necessary to develop and implement a data curation program in an academic library; engaging with exercises to facilitate discussion of varied approaches to implementing data curation programs; understanding components of a repository development plan and best practices for carrying out data curation activities. To address these objectives, we brought GSLIS faculty and staff together with expert presenters from several other organizations. Presenters came from Johns Hopkins University Libraries, the Purdue University Libraries, the University of Illinois Libraries, and the National Snow and Ice Data Center. While the focus was predominantly on scientific data, general topics covered at the Institute included day-to-day digital preservation; technical aspects of data repository systems; appraisal and selection of digital data; and resource requirements for managing a data curation program in the library organization. In addition, we held a panel presentation that included librarians engaged in data
curation activities along with domain scientists to talk about current data management and curation needs and experiences.

Our panel will include one of the institute organizers and three participants who are involved in data curation activities at their home institutions. Panelists will discuss the need for continuing professional development in data curation and the workshop as a model for preparing practicing librarians to engage with e-Research, Cyberinfrastructure, and data curation activities. Following an introduction on the Institute, the other panelists will present on their experiences with the Institute in relation to the current and ongoing data curation activities at their own universities. We intend for this panel to include time for discussion with the audience on the needs for and approaches to continuing education and professional development in data curation and stewardship in the academic library.

2. PANELISTS
Jan Cheetham is a scientist and academic technology consultant at University of Wisconsin-Madison’s central IT unit, the Division of Information Technology, where she collaborates with librarians, researchers, and IT experts to help shape campus cyberinfrastructures for data curation and preservation.

Melissa Cragin is the Project Coordinator for the Data Curation Education Program at the Graduate School of Library and Information Science at UIUC, and a doctoral candidate conducting research on shared scientific data collections and implications for library-based data curation.

Leslie M. Delserone, the Agriculture Librarian at the University of Minnesota Libraries, is a former plant pathologist and graduate of the IMLS-funded “Program for University Librarians in the Sciences” from the University of Iowa School of Library and Information Science.

Marianne Stowell Bracke, Agricultural Sciences Information Specialist at Purdue University, is involved in creating metadata schemas for several data archiving projects and is working to create dialogues within the agricultural disciplines on data curation issues.

3. REFERENCES

ABSTRACT
The Open Archives Initiative (OAI) has recently created the Object Reuse and Exchange (ORE) project that defines Resource Maps (ReMs) for describing aggregations of web resources. These aggregations are susceptible to many of the same preservation challenges that face other web resources. In this paper, we investigate how the aggregations of web resources can be preserved outside of the typical repository environment and instead rely on the thousands of interactive users in the web community and the Web Infrastructure (the collection of web archives, search engines, and personal archiving services) to facilitate preservation. Inspired by Web 2.0 services such as digg, del.icio.us, and Yahoo! Buzz, we have developed a lightweight system called ReMember that attempts to harness the collective abilities of the web community for preservation purposes instead of solely placing the burden of curatorial responsibilities on a small number of experts.

Categories and Subject Descriptors
H.3.5 [Information Storage and Retrieval]: Online Information Services—Web-based services; H.3.7 [Information Storage and Retrieval]: Digital Libraries—Collection

General Terms
Design, Experimentation, Human Factors

Keywords
digital preservation, OAI, resource maps, web resources, web curation

*This work performed while Dr. McCown was working with the Digital Library Research & Prototyping Team at the Los Alamos National Laboratory.

1. INTRODUCTION
The Web continues to be one of the most useful constructs to disseminate information, enable mass communication, and document our lives. There are, however, two notable challenges, among many, that confront the Web. The first challenge is curatorial. The Web is very difficult to curate because of its sheer size and distributed nature, its lack of editorial control and ephemeral qualities. Web pages that are here today are often gone tomorrow, and links that were once valid now return 404 responses or material that no longer reflects the original link creator’s intent. The transient nature of the Web has been addressed by a number of parties: web archives like the Internet Archive\(^1\) store historic snapshots of the Web, search engines like Google make temporarily inaccessible pages available from their caches, and personal archiving tools like Spurl\(^2\) and WebCite\(^3\) let users archive individual web pages for viewing at a later time. Although none of these strategies in isolation are completely effective at fending off link rot, the combined efforts of these services, what we call the Web Infrastructure (WI), provides a layer of preservation which adequately protects a massive number of web resources [12].

The second challenge facing the Web is organizational in nature. The Web has previously lacked widely accepted standards to group distinct web resources together into a whole. There are many times when a resource, like an online book, academic publication, or news article, is composed of separate web pages or other web-accessible resources. Although it is usually easy for humans to determine the boundaries of such aggregate resources, it is problematic for an automated agent to do the same [2].

In response to this challenge, the Open Archives Initiative (OAI) has created the Object Reuse and Exchange (ORE) project which provides standards for defining and discovering aggregations of web resources [6]. An aggregation (sometimes called a compound information object [6] or compound document [2]) may be composed of text, video, images, and any number of web-accessible, URI-identified resources. For example, a scholarly publication may consist of an HTML “splash page” along with versions of the paper in PDF and PostScript format, a video slideshow; and the raw data used to perform the related research. An aggregation documenting a special event like 9-11 could be composed of images,

\(^{1}\)http://www.archive.org/
\(^{2}\)http://www.spurl.net/
\(^{3}\)http://www.webcitation.org/
The ORE Data Model introduces the concept of a Resource Map (ReM), a web resource that describes an aggregation. ReMs act as an organizational unit, defining the boundaries of an aggregation and indicating the relationships between the aggregated resources. Like their aggregated resources, ReMs have their own URIs. They may be housed in an institutional or academic repository like arXiv\(^4\) where they may receive a high degree of monitoring by administrators. Others may exist outside the repository where they may be maintained by any number of individuals.

Unfortunately, whether ReMs are maintained inside the walls of a repository or outside in the wild, they may eventually fall prey to neglect. ReMs share many of the same preservation difficulties as other web resources: ReMs may change over time, move to different URLs, or disappear completely from the Web. However, they also present additional challenges because the resources they aggregate may also change, move to different URLs, or disappear. This added dimension suggests ReMs may require more curatorial attention than other web resources.

In this paper, we explore some strategies that can free the ReM creator from the burden of full curatorial responsibilities and instead distribute or democratize the workload to the masses. Inspired by Web 2.0 sites that rely on the public for producing and maintaining content, we have developed a system called ReMember which leverages the distributed efforts of the public who interact with web archives, search engines, and personal archiving services (the Web Infrastructure) to maintain the integrity and accuracy of ReMs. Employing a small number of experts to provide equivalent curatorial services would be prohibitively expensive and would not scale to the Web. But by distributing the effort to the public, we believe the small, contributed efforts of many in conjunction with the WI will allow us to curate ReMs on the scale of the Web.

2. BACKGROUND

2.1 Object Resource and Exchange

As mentioned earlier, humans can easily determine the boundaries of an aggregation, but it is very difficult for a machine to do the same. Sharing the Semantic Web’s goal of enabling a machine-readable Web and the Linked Data vision of connecting disparate datasets, the ORE project aims to create standards that allow aggregations of web resources to be defined and discovered [7, 8, 16]. These aggregations are conceptual resources which are made concrete by Resource Maps (ReMs). ReMs enumerate the aggregated resources (ARs) that make up an aggregation and include descriptive metadata about each AR. Aggregations and ReMs have distinct URIs, but dereferencing an aggregation’s URI will lead to the authoritative (or trusted) ReM that describes it.

An example ReM is shown in Figure 1 where ReM-1 is the URI identifying the ReM, and A-1 is the URI identifying the aggregation. The aggregation contains three aggregated resources (AR-1, AR-2, and AR-3), and RDF triples are used to describe the relationships between the ReM, aggregation, and ARs.

ReMs may be serialized in a number of formats like RDF/XML and RDFa, but the simplest format is the Atom Syndication Format [13]. Atom is a popular syndication format for blogs, but increasingly it has been used for other purposes like the Google Data API\(^5\).

An example ReM\(^6\) for an arXiv e-print is shown in Figure 2. This simple example shows two ARs, an HTML “splash page” and a PDF, that, together with other ARs not shown in the example, constitute the e-print resource. The ReM itself was last updated on 2007-10-10, but an updated timestamp of 2006-05-31 was used for the ARs; this later timestamp may reflect when the ARs were last modified or when the Atom entry was last modified. Both ARs have link elements which indicate the URLs of the resources.

ReMs may be created by anyone. They may be discovered by humans and bots by following a link to them. An HTML resource that is aggregated by a ReM may also contain a link element which points to the ReM; non-HTML ARs can make use of the HTTP link response header which performs

---

\(^4\)http://arxiv.org/

\(^5\)http://code.google.com/apis/gdata/overview.html

\(^6\)The example uses version 0.9 of ORE which has recently been replaced by version 1.0. Although the Atom serialization in version 1.0 is significantly different, it can easily be implemented in ReMember without any impact on its functionality.
the same function. Batch discovery methods like SiteMaps and OAI-PMH may also be used to discover ReMs. More technical details of ReM Atom serialization and other ORE standards are available on the OAI-ORE website [7].

2.2 Web Infrastructure

The Web Infrastructure (WI) is the collective activities of web archives (e.g., Internet Archive), search engines (e.g., Google, Live Search, and Yahoo), personal archiving tools (e.g., Spurl, Hanzo:web, and WebCite), and research projects (e.g., CiteSeerX and NSDL) that refresh and migrate large amounts of web content as by-products of their primary services [12]. The WI can be used as a passive service for a number of preservation functions. For example, websites that have been lost without backups can be reconstructed from the WI using Warrick [10], and web resources that move from one URI to another can be relocated using Opal [4].

Figure 3 illustrates how the WI has captured multiple versions of the ORE home page. The upper-left screenshot shows Google’s cached copy of the page as they crawled it on 2008-07-12. The middle screenshot shows WebCite’s archived version of the web page from 2008-07-21 (as initiated by one of the authors), and the bottom screenshot shows multiple versions of the same page available from the Internet Archive (IA) from 2006-11-06 to 2007-08-21. Unfortunately, IA has a 6-12 month lag in their archive, so they do not have any copies available from 2008.

There are some notable differences between various members of the WI. Search engines usually have the most up-to-date and widest breadth of resources available from their caches because of the competitive nature of web search and huge investments in web crawling infrastructure [9]. However, when search engines discover that a web resource has been changed, they discard the old version of the resource for the new one. They also migrate textual resources like PDF and Microsoft Word into HTML pages that lose their formatting and embedded images.

Web archives like IA also rely primarily on web crawling to discover web resources. IA keeps old versions of resources in their original format, but as stated before, they are slow.

Figure 1: An example Resource Map, borrowed from [7].

Figure 3: Copies of the ORE web page at http://www.openarchives.org/ore/ stored in Google’s cache, the WebCite archive, and the Internet Archive.

3. CLIENT-ASSISTED PRESERVATION

3.1 Overview

Several of the WI members like the Internet Archive and commercial search engines rely primarily on web crawling to populate their repositories. This automated “pull” methodology, illustrated on the top pane of Figure 4, works well in terms of finding a large number of web resources, but some may be missed (e.g., resources that require too many hops from the root page, resources for which links may not be found, and “unpopular” resources).

With the introduction of ORE (middle of Figure 4), ReMs (red dots) delineate the aggregated resources on the Web (for simplicity, aggregations and aggregations’ ReMs are shown as a single unit). The WI may archive and cache some ReMs and ARs as they do any web-accessible resource. In previous work, we showed how IA could archive evolving ReMs and their evolving ARs without any architectural changes [15].

By introducing our ReMember system, we hope to enable millions of Web users to perform a small amount of curatorial work in keeping ReMs that change over time current by ensuring that all AR URIs resolve at various points in time to the correct content. Individuals will push ReMs and ARs into the WI, allowing for more complete and accurate coverage than what is now possible (bottom of Figure 4).

3.2 Architecture

In order to harness the curatorial power of the masses, we have constructed the ReMember prototype which adheres to several important design goals:

1. Resource producers should easily enable inclusion of their resources and ReMs into the system.
2. The system should rely on the WI for storage since we do not personally have sufficient storage capacity for potentially millions of resources.
3. The system should not assume any WI member will always be accessible. Members of the WI may come and go, and therefore copies of the resources should be spread throughout the WI.
4. The system should help users relocate missing resources by maintaining a small fingerprint that might help identify the resource.
5. Changes to ReMs should be logged over time to allow for rollback operations.

ReMember is a lightweight system that attempts to put as little demand as possible on resource producers and consumers. In order for resource producers or maintainers to mark their ARs and ReMs for inclusion in ReMember, they may insert an HTML snippet into the bottom of their HTML page that produces a “Preserve this Object” link. When a user who is viewing the page clicks on the link, the user will be prompted to preserve the ReM and its ARs. If AR maintainers are unable or unwilling to add an HTML snippet to their resources, users may still curate ReMs and their ARs by use of a browser plug-in (to be implemented) that automatically discovers ReMs and facilitates submitting them to ReMember. Individuals may also submit ReMs to ReMember directly using its web interface.

In the future, we envision a del.icio.us-like interface that displays the ReMs that users are most frequently preserving. But unlike del.icio.us, we may be more interested in showing users the resources that are not being preserved, the unpopular ReMs, since these resources are in most need of the community’s attention.

The architectural and process overview of ReMember is shown in Figure 5. When a user who accesses an aggregated resource clicks on the “Preserve this Object” link, the AR’s ReM URL is submitted to ReMember. If the ReM has never been seen by ReMember, it will immediately push a copy of the ReM and each AR (obtained from the ReM) to the WI via a personal archiving service (WebCite). ReMember will also create a lexical signature (five words that could be used to uniquely identify the resource when searching the Web [14]), store the ReM in its wiki (for version control), and store a thumbnail image snapshot of each AR. The lexical signature and thumbnail, together with the URL and ReM AR metadata (title, author, description, etc.), will act as a fingerprint for an AR in case the WI were to lose its copies or if the user needed to find an AR that moved to a different URL. If the user is willing, he/she will also submit the URL to several search engines if they have not yet indexed the
AR (this usually requires the user to solve a CAPTCHA and thus cannot be fully automated).

Subsequent accessing of the ReM in ReMember will prompt the user to correct any broken links to missing ARs. This involves showing the user any older copies of the AR that may be found in the WI using the AR’s old URL. The user can also search the Web for the new location of the missing AR using the metadata from the ReM and the lexical signature. The user will also be prompted to examine any ARs that have changed since the last time they were archived in the WI; significant changes warrant re-archiving, obtaining new lexical signatures, and creating a new thumbnail. Any changes made to the ReM are archived to the WI, and the changes are noted in the wiki. By ensuring that the ReM is valid each time a user visits ReMember, we may allow users to view older versions of the ReM and its associated ARs at various points in time by pointing to archived versions of ARs in the WI.

The following summarizes the data being stored in ReMember and the WI:

**Stored in ReMember**

(a) ReM at time $t_i$ (document)
(b) For each AR$_j$ in ReM at $t_i$ (URI)
   1. Metadata (title, author, etc.)
   2. Lexical signature
   3. Image thumbnail
   4. URI of AR$_j$ in WI

**Stored in the WI**

(a) ReM at $t_i$ (document)
(b) AR$_j$ at $t_i$ (document)

### 3.3 Possible Scenarios

A ReM and its ARs might exhibit a variety of changes over their lifetimes as illustrated in Figure 6. The vertical lines at $t_1$, $t_2$, and $t_3$ represent users accessing ReMember at various times and curating the ReM. The diagram is not exhaustive, but the most common events are accounted for. Figure 6 shows six ARs (1-6) that are created before the ReM. The ARs in Figure 6 are added to the ReM by the ReM creator at creation time. AR7, which is not created until some time later, is also added to the ReM before time $t_3$ by the ReM’s maintainer.

Figure 6 shows AR3 moving from one URI to another before anyone has had a chance to access the resources through ReMember. So when one of the resources is accessed at time $t_1$, all the ARs can be archived except AR3; a user is required to investigate where the resource has moved. At this point, there are two components that ReMember can use to assist the user: the AR’s previous URI and the AR’s metadata. ReMember will aid the user in finding the resource by first examining the WI for copies of the resource using the old URI (a search engine’s cache or IA). Having access to old copies may help the user locate the new URI of the AR with the aid of a WI search engine like Google. Even if old copies are not found, the metadata can be used as a query in a web search. Once the user finds the resource’s new URL, the ReM is updated with the new information, the ReM is archived in the WI, and the changes are logged on ReMember’s wiki.

Figure 6 shows that AR4 also changes URIs, but this takes place after ReMember has had the chance to create a lexical signature and thumbnail snapshot of the resource. This information, along with the metadata stored in the ReM, can be used to help the user at time $t_2$ to relocate the resource. AR2 has been completely removed from the Web between $t_1$ and $t_2$, so the user will not be able to find a new copy of it at $t_2$. This requires the user to flag the resource as being no longer accessible on the Web. The same decision will need to be made for AR3 at $t_3$ when the URI still resolves but points to the wrong content, and the AR is not available at any URI except at WebCite. AR1 was removed from the
ReM outside of the ReMember system sometime between \( t_2 \) and \( t_3 \), but user intervention is not required; ReMember need only archive the new ReM and capture the changes in its wiki.

AR1 and AR5 undergo some degree of change between \( t_1 \) and \( t_2 \), both of which require a user to decide if the change is significant enough to warrant re-archiving the AR or finding a suitable replacement AR. AR1 experiences only a minor change, like a change in the date or an advertisement or maybe a minor layout adjustment. AR5 undergoes a significant change, like an update to a blog entry or new version of an academic paper. Although heuristics can be devised to determine the degree of change, we believe a human (assisted by the heuristics) is more likely to make the best curatorial decisions.

4. EXAMPLE: ACADEMIC BIBLIOGRAPHY

To illustrate how ReMember might be used in a real world scenario, we created a ReM based on a bibliography found online\(^8\) that points to twenty-six online papers about digital preservation (the papers were housed on multiple websites). The web page serves as a human-readable aggregation, but a machine would have difficulty determining which links were to be included in the aggregation and which simply pointed to related websites. We added the splash page and each of the bibliographic entries to the ReM as aggregated resources. The title of the papers and authors were entered as metadata for each AR.

When accessing the newly created ReM for the first time in ReMember, the ReM is pushed to the WI (WebCite) and to the wiki. Each AR is downloaded, and a screenshot thumbnail and lexical signature is created for those ARs that are successfully downloaded. The ARs are also pushed to WebCite.

A screenshot of ReMember is shown in Figure 7 as the user would see it when accessing the ReM. Each AR has a thumbnail image shown on the left with its accompanying title, updated timestamp, author, and any other descriptive metadata.

As indicated in the screenshot, the splash page (the first AR) does not need any user attention since this is the first time ReMember has seen this AR, and it was successfully accessed. The second AR, however, returned a 404 response when ReMember attempted to download it, so the user’s assistance is required. When the user clicks on “Needs attention,” a new browser window will appear which will first show any copies of the AR that the WI may have. Since the resource has been missing from the Web for some time, the search engine caches no longer have a copy, but IA has a version from 2007-01-06. The user could update the ReM to indicate the resource cannot be found on the live Web. The ReM’s changes are logged in the wiki. A screenshot of the wiki is shown in Figure 8 after making several changes to the ReM. Users may view this wiki at any time and can rollback any erroneous modifications that may have been made.

The next time the same ReM is accessed in ReMember, the live ReM will be checked to see if it has undergone any changes since the last time it was archived. Each AR is also accessed and compared with its archived version in WebCite.

\(^8\)http://www.chin.gc.ca/English/Digital_Content/Digital_Preservation/bibliography.html

Figure 7: Screenshot of ReMember.

Figure 8: Screenshot of ReMember’s wiki for ReM version control.
change of some sort, ReMember will request the user’s attention for this AR, and the user may decide that the newer version should be archived. ReMember will push a copy of the updated article to WebCite, compute a new lexical signature, and take a new snapshot.

ReMember also allows users to view a timeline of changes that occur to ARs using MIT’s Simile Timeline\(^9\) widget. Figure 9 shows a timeline for the ARs in this case study’s ReM. All the resources were archived for the first time on Aug 7, and several ReMs experienced significant and insignificant changes on different days; some moved to new URLs or went missing on the live Web. This visualization helps users see which ARs are the most volatile over time. While ReMember does not attempt to show explicit differences between each version of an AR, other projects like the Past Web Browser do [5].

Appendix A shows several other example ReMs being curated in ReMember.

5. ONGOING WORK AND CONCLUSIONS

Link rot has been a continual adversary of the Web, and a number of solutions have been offered to combat the problem (e.g., [1, 4, 11]). Our system does not replace such systems but augments them by attempting to harness the abilities of the web community to curate ReMs when automated processes are not enough.

As ReMs become more prevalent on the Web, we hope ReMember will be embraced by a community of individuals who desire to keep ReMs on particular topics accurate, just as Wikipedia has been embraced by a large community to curate a large number of articles on a variety of topics. Like Wikipedia, our system will likely be targeted by spammers, and it remains to be seen what editorial controls or techniques will be required to fight mischievous alterations.

We are also investigating how the community could take ownership of a ReM and add and delete ARs from it. We believe that the community would take more interest in curating ReMs if they could personally enhance its usefulness like one might enhance a Wikipedia article. Again, spam issues will likely be a significant challenge.

6. ACKNOWLEDGMENTS

This research is supported in part by the Andrew Mellon Foundation. We thank members of the LANL Digital Library Research & Prototyping Team, especially Ryan Chute, for their technical assistance in developing the ReMember prototype.

7. REFERENCES


---

\(^9\)http://code.google.com/p/simile-widgets/wiki/Timeline
APPENDIX

A. ADDITIONAL EXAMPLES

Some additional examples are included here to show different types of ReMs being curated with ReMember.

Figure 10 shows a ReM that points to a number of web pages about the Denver Broncos football team. The second AR needs attention because its text has changed significantly since the last time the AR was curated. The third AR is still being checked; ReMember checks resources asynchronously since some web servers may respond slowly to requests.

A ReM that points to various news reports, images, and videos (all from CNET.com) of the proposed Microsoft-Yahoo merger is shown in Figure 11. Finally, Figure 12 shows the ReM from the arXiv e-print example of Figure 2.
Digital Curation and the Citizen Archivist

Richard J. Cox
School of Information Sciences
University of Pittsburgh
Pittsburgh, PA 15260 USA
rcox@mail.sis.pitt.edu

ABSTRACT
The increasing array and power of personal digital recordkeeping systems promises both to make it more difficult for established archives to acquire personal and family archives and less likely that individuals might wish to donate personal and family digital archives to archives, libraries, museums, and other institutions serving as documentary repositories. This paper provides a conceptual argument for how projects such as the Digital Curation one ought to consider developing spinoffs for archivists training private citizens how to preserve, manage, and use digital personal and family archives. Rethinking how we approach the public, which will increasingly face difficult challenges in caring for their digital archives, also brings with it substantial promise in informing them about the nature and importance of the archival mission. Can the Digital Curation project provide tools that can be used for working with the public?

Keywords
Archival advocacy, blogs and blogging, citizen archivists, collecting, diaries, digital curation, electronic mail, ephemera, family archives, Internet, personal archives, personal computing, photography, recordkeeping, scrapbooks, World Wide Web.

1. INTRODUCTION
Much has transpired with personal information technologies in the past two decades, suggesting that personal archiving, from websites and blogs to digital photograph albums and scrapbooks, is already a prominent feature of our society, as well as a prominent preservation issue.

The present interest in personal archiving represents a major new opportunity for archivists to re-imagine and better communicate their mission in society by aiding individuals who have already developed some interest in the archival enterprise. We need archivists to develop innovative publications, Web sites, and other training materials to assist the public. Archivists now have the opportunity to connect with a growing portion of the public looking for advice about preserving personal and family documents. Archivists may need to alter their mission and priorities, but the possible results may be unprecedented in terms of gaining public support and understanding.

My question in this paper is whether the new interest in digital curation and in developing a multi-faceted curriculum for educating the next generation of archivists has some value for working with the public. This paper is conceptual. It is difficult to pin down the present digital curation effort because it is in development. However, given the purpose of this project, it is possible to make some proposals for why we need to develop some spin-offs to be used for equipping citizen archivists to work with their own personal and family archives, increasingly created or stored as digital objects.

2. ARCHIVAL APPRAISAL AND INDIVIDUAL COLLECTING
Collecting is a basic human instinct and books and articles pour forth regularly about the history and meaning of collecting. Personal collecting can seem quirky or frivolous, but it always reveals some deeper inner meaning to life’s purpose. Traditional and official archival repositories have long operated on the basis of collecting personal and family papers, juxtaposing their public, institutional role against that of individual, private collecting. The nature of this contrast has been exploded by the advent of the digital era, complicating institutional collecting and empowering individual archiving, both possibly at the expense of the work of future historians and other researchers (or maybe not, instead just redirecting researchers to the World Wide Web as an archival repository) (as the number of practical manuals on the topic grows, such as Brown) [1].

One of the continuing promises of our present digital age is the idea that individuals will be able to save every scrap of information about their lives and families and call them forth effortlessly and seamlessly whenever needed. No one will deny that this is an intriguing prospect, or that it is an engaging topic to reflect on. Designers and researchers exploring the MyLifeBits software, designed by Gordon Bell, certainly reveal why digital personal archives can be so captivating. Reviewing the advances in cheap storage, desktop search tools, and metadata development, these software advocates describe a database supporting the range of personal documents each of us is likely to produce in our lifetime [2]. What we hear is the mantra, save everything. Then when we examine the uses of FaceBook and MySpace, we hear another refrain, display most everything publicly (with some personal control and restrictions).

3. PERSONAL AND FAMILY ARCHIVES
Whether we consider the use of personal information technology or traditional technologies involving ink, paper, and leather bound notebooks, there is little question that the interest in personal and
family archiving is growing. We are witnessing an upsurge in diary writing, journaling, and calligraphy – perhaps part of a reaction to the bits and bytes of the digital world [3]. Self-help publications appear regularly to assist individuals in taking up these pastimes. Attics, basements, and garages, even in the heralded era of cyberspace, often still contain boxes and bags of old family papers. As time passes, however, the kinds of documents we discover in these spaces seem more and more foreign to us.

It is important for records professionals to remember the emotions that might be associated with even the most mundane looking document. And this is certainly the case when we consider personal and family papers, especially in their traditional forms. Sometimes it is easy for archivists to lament the times that they have been forced to watch and listen as someone carefully pulls out documents one at a time and tells a story about each one. Yet, this very human response to interacting with the archival documentation giving them meaning and placement in our age ought to tell archivists something about the value of their work.

While one can read transcripts of such documents on the World Wide Web or their digital counterparts by contemporary writers, the experience between this kind of reading and possessing a physical document is not quite the same. In the past, while there has always been tension between private and public (institutional) collectors, it has been the institutional collectors – archives, libraries, museums, and historic sites – that have won out. In the future, there may be less certainty about this, especially as so many personal papers are digitally born and pose challenges to the public archives. The good news is, however, many private citizens care as passionately about the documents as do the institutional repositories.

People may be increasingly reluctant to turn over their original documents to archives, libraries, and other institutions because of sentimental and emotional meanings not being afforded by the new and emerging digital documents. Before the computer, of course, people worried about the increasing number of technologies encouraging oral rather than written transmission, and their impact on written records. However, there may be less to learn looking backwards because of the remarkable advances in affordable information technologies. Whatever perspective archivists might assume about shaping the documentary heritage through planned appraisal approaches, private individuals will continue to save their own personal and family archives and, different than what has occurred in the past, we might see these documents not hidden away but visibly posted on the Web. Archivists must explain and advise about the basic tasks necessary to maintain archival documentation, requiring new depths of technical and other knowledge, such as intellectual property and personal privacy.

4. ARCHIVING AND EXPERTISE

The world is changing, at least in how it views expertise. It may seem ironic, but the many preservation challenges posed by digital technologies also suggest more empowerment to individuals to administer their own personal papers. Individuals who are not archives and records management professionals may be interested in developing their own expertise to administer their part of the larger documentary heritage (even though solving the challenges cannot occur without some expertise or more resources, public policies, and laws and regulations). We can learn a lesson from the historical development of writing: computers will become so essential that they will become available to all and in ways that are seamless and painless, just as formal scripts were supplanted by the less formal cursive for everyday recording and communicating [4].

It is worth noting that for all the claims made about digital writing and the World Wide Web that the prospects for preservation seem all the more dim. Those working within the digital curation movement reflect this as their reason for working on technical models and methods. Martin Halbert writes, “CMOs [cultural memory organizations] hold virtually innumerable archives of idiosyncratic material that are rapidly being digitized in local initiatives. This digital content has important long-term value for both research and cultural identity purposes. But CMO professionals frequently lack effective’, scalable DP infrastructures. This lack of access to effective means for long term preservation of digital content is aggravated by a lack of consensus on DP issues and professional roles and responsibilities” [5]. This perspective also helps us to understand that the reason people cling to old family papers or try to administer their own records may have little to do with some noble societal cause, but everything to do with personal interests, curiosity, and self-identity – or recognition of the fear of the loss of digital stuff that they read and hear about. Archivists should want to nurture such interests, not diminish them, because they support the mission to preserve our documentary heritage.

Is it no wonder then that even as every home acquires a personal computer the shelves in the home are also being filled with leather-bound, acid-free journals for diaries and commonplace books? The only differences between these earlier revolutions and the present computer era is the amount of time involved; centuries have shrunk to decades, decades to years, years to months (it is often the compression of time in new technical developments that is identified as the main attributes of the present information age). One can practically use and cherish an old fountain pen or camera for decades, but every few years we need to replace our computers (even after they have been upgraded numerous times) [6].

5. THE CHANGING WORLD OF PERSONAL RECORDKEEPING

We have new challenges involving personal recordkeeping. The theft of laptops, identity theft, ownership and responsibility for personal medical records, shifting and confusing notions of personal privacy, and digital documents replacing paper forms and posing new maintenance challenges have all transformed the notion of personal and family papers. How well we do with these personal papers is critical to our identity. We are surrounded by documents marking the activities of our lives, the history of our families, and the unrelenting passage of time. Bills to be paid pile up on our desks at home. Papers from our workplaces can usually be found nearby, or, if we are efficient and organized, in our briefcases, packed and ready to be consulted and worked on. We save certain documents, an interesting letter from a family member or an annotated greeting card from a friend, as mementos of important events in our lives. We assiduously maintain our financial records, carefully organized by accounts and functions, and usually reflecting our sense of how we will tackle the
unpleasant annual chore of filing our income tax statements. Photographs, diplomas, and certificates of awards are framed and decorate parts of our houses and offices. Sometimes we use the most routine documents, such as checks, to recreate a life [7].

These witnesses have been with us for a very long time. The impulse to record extends back tens of thousands of years and is seen in the cave paintings, decorated objects, and other material culture remains left us by early humans. Writing systems are, of course, much more recent innovations, but they tell us remarkable things about ancient societies, including what they ate, how they traded with each other, who the rulers were, evil acts perpetrated on people, stories of miracles and great beneficences, natural disasters, wars, what people wore, and how they built residences and public buildings. Although we view our own age as the time when great quantities of information are created, maintained, used, and abused, such recording is endemic to human nature – and that all eras are eligible to be termed “information” ages [8].

Even if one could argue that the impulse to document our activities is not part of our human nature, it is hard to argue that the sources driving writing and recording were not connected to the most basic of human functions. The most mundane of all recordkeeping, tracking financial transactions, is probably the oldest records system known to us. The most common financial record is our checking book. Although many banks have ceased sending cancelled checks back to the customer, there are billions of these checks floating about.

Letter writing has been around since the ancient world, and it shows little signs of disappearing. Most Americans take for granted the daily arrival of mail at their doorsteps, even as their increasing use of electronic mail has affected how, when, and why they choose to write a letter, affix a stamp, and drop it into a corner mailbox. Somehow, however, the letter continues to hang in there. People certainly use short-cuts in our faster-paced world, such as writing long messages in pre-fabricated greeting cards or postcards or mass-producing on word processors what appears to be personalized letters, but the function and allure of the letter remains intact. The major change in letter writing has come in the form of electronic mail, and electronic mail is one of the primary features of our modern networked society, where one can communicate nearly instantly with others where it used to take days or weeks before. The use of digital letter writing poses problems regarding the maintenance of a personal archive.

Most of us, at least those of us at a socio-economic level where we own substantial property, also are cognizant of the need to maintain property records as a form of protection and a manifestation of the responsibility that comes with property. Like financial records, property records extend back to antiquity. As commerce and government developed, the need for documenting the ownership of land and houses emerged, and the nature of marking physical features on the landscape soon proved unreliable as the ownership of property became more complicated (although the earliest records, predating scientific systems of surveying and mapping, often documented immense amounts of these features and that of oral tradition as well) [9]. All those old metal document boxes we find in antique stores and flea markets are testimony to the fact that people have been maintaining property and other vital records in safe places for a very long time; examples of the predecessors of such document boxes date back to ancient society, and the storage devices and the function they represent provide an easy to comprehend link between what organizations and governments do with their records and why we manage our personal papers.

The intensely practical merits of financial and property records and personal correspondence ought not to overshadow the equally intensive personal needs to create and maintain records. One historian’s study of Abigail Adams, the wife of John Adams, describes her letter writing as having a “therapeutic function.” “Abigail had the rare capacity to express her grief, anger, and fear in words on paper. In doing so, she also helped to raise her own spirits. By transferring her emotions to paper and then mailing the letter, she banished her unhappy temper.” Abigail’s letter writing was a means for her to “unconsciously transform the raw experience of her daily observations into a strongly formulated system of values. Writing to the folks at home about the strange scenes encountered in her travels confirmed their reality in her own mind as well” [10]. Whether one is using quill, ink, and paper or tapping away on a computer keyboard, the therapeutic aspects of personal recordkeeping have not lessened.

6. HUMAN IMPULSES AND PERSONAL ARCHIVING

We have a great need to consult regularly many of our personal archives. We re-examine property records when we are contemplating selling them or refinancing them. We inventory when we are working on a will or updating one. We check financial records when we believe we have been over-charged for a purchase. We pour over old family papers when we need a photograph of an ancestor or a remembrance of a past event to be used in the production of a greeting card, wedding invitation, or renewal of wedding vows. These human impulses to record events and then to save records are reflected in legal matters, societal customs and traditions reflected both in the actual forms of the records and other sources such as etiquette manuals, and in our display of older records – framed and strategically situated – in our houses, workplaces, and our wallets and computers.

New technologies have changed the process but not necessarily the aim of how we interact with our personal and family archives. Digital photography has a more tenuous relationship to reality than that of earlier photographic forms. Now a photograph is information and does not become an image until called up and tinkered with, exaggerating all of the earlier debates about just what a photograph’s image is – art or reality, for example. Critics, historians, and other scholars long ago abandoned thinking of any photographic image as just a frozen moment in time and space, adopting far more complicated concepts of what the image is, but a digital photograph seems more complex by many orders of magnitude, mainly because it is so much more malleable [11].

Iconography is the study of images and their symbolic role in our lives, institutions, and society. And, as such, it also speaks to the use of documents in displays in our homes and offices meant to interpret our life. Often displayed with the care of a museum exhibit, these spaces tell us much about how and why people want to preserve at least a portion of their private archives. The public display of documents is a means by which we connect with the past. The mere retention and management of our personal archives suggests this purpose as well, but such administration is often done behind the scenes, with records neatly stored in boxes.
and folders, on disks, and on personal computers. The public display employs a much more selective process of interpretation, whereby we assemble key documents – sometimes selected as much for their aesthetic value as for their evidence – to portray a certain image or to assume a particular identity.

We can understand more of this role of personal archives if we understood why we carry certain records with us. The photographs, receipts, credit cards, licenses, and membership cards we transport with us every day in our wallets, purses, and briefcases speak loudly about us as citizens of the world. Some of these cards, such as a driver’s license, we carry with us because we are required to by government agencies. We keep these documents with us because they provide some identity for us, especially as we relate to others. And, with the aid of laptop computers, PDA’s, and cell phones, we can now carry far more personal information, some of it quite symbolic of who we are and much of it as carefully arranged and catalogued as an exhibition at a museum (or at least as good as the hallway outside the kitchen).

Recordkeeping we associate with government responsibilities as public data managers, such as vital family records – like death, birth, marriage, and baptismal documents – and a large array of licenses (from hunting to driving, passports and professional certification) have seemed to be with us nearly forever. Many people associate such vital records with government responsibilities, and, indeed, such recordkeeping did become an essential and ubiquitous aspect of government bureaucracies as social, health, and legal services expanded. Some associate such government recordkeeping as being synonymous with bureaucracy, the filling out of endless forms that every citizen experiences every time they visit a government office. Nevertheless, our effort to complete such records or to provide the information essential for completion leaves traces everywhere of us, adding to the accumulation of personal records we generate on our own volition.

Vital recordkeeping was not always a government responsibility. Non-governmental organizations, most notably churches, recorded births, deaths, marriages, and baptisms as part of their sacramental responsibilities – and it is reasonable to assume that if government had not become the official agent for this that such private groups would have continued to perform this function. Many of us have copies of these documents in our family scrapbooks or framed and hanging on the wall, especially since many of these documentary forms are in beautiful calligraphic hands and are associated with landmarks in our lives and those of our families. The rapid growth of interest in genealogy through the past century shows no signs of abating and, along with other hobbies such as scrapbooking and diary writing, suggests a continuing interest in personal archives [12]. If government was not recording so many of our activities, even with all the reasons of privacy invasions and misuse of personal data that should concern us, would we simply ramp up our own self-recording? The relationship between governmental and organizational recordkeeping and the individual impulse to develop personal archives is a complex, but quite real and useful, one.

The level of commitment we might want to invest in such personal documentation, even to the point of forging our own documentary past, can also be seen in other document forms, such as diaries. The writing of diaries has been a human activity for centuries, and nearly everyone can think of a famous one that has been published. And, perhaps, the best window into the process can be seen in the words of diarists themselves, such as in the those of Rev. Francis Kilvert, a nineteenth century English curate, quoted by Bret Lott in his book on writing: “Why do I keep this voluminous journal? I can hardly tell. Partly because life appears to me such a curious and wonderful thing that it almost seems a pity that even such a humble and uneventful life as mine should pass altogether away without some record such as this” [13]. The process of diary writing may be the quintessential act of personal recordkeeping, where the daily – or some regular occurrence – of scribbling in a bound book can become an obsession of trying to record every activity, or, at least, an interpretation of every activity.

Whatever diaries might be, they have become popular again. It seems that nearly everyone at least starts compiling a diary, although most do not sustain the process (committing to a diary over an extensive period of time has about the same success rate as dieting and New Year’s resolutions). There is a contemporary revolution in diary writing. Diaries are portrayed as the place to write down those salacious thoughts about illicit or immoral activities, a kind of protest against authority (but, also, I would argue, really intended to be read as well). Now, of course, every bookstore chain devotes an entire section to “archival” quality notebooks, beautiful and sleek fountain pens, and books about how to compile a diary (along with family histories, photograph albums, and scrapbooks). There is a public invitation to creating personal archives everywhere one looks, and it is sometimes hard to ascertain whether the vendors are driving and creating a market or whether a market has emerged all on its own. The popularity of diary writing can be seen in the growing presence, since 1995, of online diaries on the Web, with their writers searching for an audience, a connection with others.

Blogging eliminates the middle agent, removing the question of whether the diary is intended for personal use or for public consumption by going instantly to a public (one much larger than when occurs when diaries are placed in archives or even when they are published). Blogging suggests that personal archives are becoming more public, as people also put scrapbooks with family photographs and scanned images of memorabilia online for their family and friends. As one thinks about their personal recordkeeping, it is not difficult to imagine a clear or precise role for the diary. It is the backbone of a personal archive, providing the basic outline of a life and its activities, a frame of interpretation for other records.

As photography became cheaper, more portable, and more adaptable for a wider range of activities, individuals and families took to photography as a means of documenting all facets of their lives. Photographs of travels and tourism became a normal personal pastime for the use of the camera. Family events, such as picnics, baptisms, and reunions, were all documented less formally than before. Informal images, people reading or talking over dinner, proliferated and filled scrapbooks and document boxes. A fuller sense of the personal archive became possible because of the advent of photography, and it is safe to say that newer technical developments, such as Polaroid instant photography and portable digital cameras, added to the possibilities of a richer documentary foundation (although posing new preservation problems).
Another essential part of any personal archives is the array of certificates one gathers over the years. We accumulate diplomas, award citations, and certificates indicating the completion of a special course or program through the years, with the largest clumps coming in our earliest and latter years. Our parents dutifully keep those certificates we receive in grade, middle, and high school, and we generally start holding onto them when we enter our college years. Most of these documents are intended to be framed and displayed, much like what we do with photographs. These documents are markers of personal progress, expertise, and authority, and they are most often displayed in public spaces such as our offices.

Such records exude symbolic value and provide a connection between modern records and their ancient antecedents. Documents such as modern diplomas have little, if any, legal value as they are issued merely for decorative purposes, mimicking older parchments and more ancient forms of records. They speak eloquently to the symbolic role of archives, a role whereby the documents take on more of a cultural rather evidential purpose. They provide a source of identity, prestige, and status, especially as they are usually displayed before us, usually hung in a place where they will be best seen. These documents manifest much of the overall symbolic value of our personal archives, including those parts of the archives that we carry around with us.

Certificates and diplomas often wind up in scrapbooks, another essential aspect of our personal archives, and now a modern multi-billion dollar a year industry with millions of people creating scrapbooks in one recent year, some using software enabling digital scrapbooks to be created as well. Scrapbooks are what many people think of if you ask them about their family archives, and, in many ways, they are another form of symbolic personal archives, with their arrangements creating order and meaning. Assembling a scrapbook is the amateur’s approach to the classification and ordering of information done by librarians and archivists, portraying what their families have been up to through the decades. Individuals select documents and enhance their value as they sort through boxes and file cabinets jammed with records, ephemera, and artifacts and shape to their own and family’s history by arranging the materials in volumes with narratives and interpretations. Menus, postcards, ticket stubs, letters, receipts, and other items serve as mementos of favorite or benchmark events, allowing individuals to construct a narrative of their past, as deliberatively as others write diaries.

7. TRACES OF OURSELVES

Personal archives can be viewed as crucial aspects for knowing about ourselves, our families, and our times, as important as licenses and memberships enabling us to function on a daily basis. There is some security in being surrounded by evidence of our lives and families. And there ought to be a feeling of insecurity when we lack such knowledge, equivalent to being illiterate. When we lack the right records, and hence the necessary information, we weaken our ability to read and cope with the world. Some of our fascination with old manuscripts, photographs, and other original documents derives from romanticized aspects of creating and interacting with such materials, but there are other utilitarian aspects motivating our personal recordkeeping.

The next time you are walking on a busy street or traveling through an airport, observe what people are carrying. Nearly everyone is laden with a briefcase, a large purse, or a backpack of some variety. Some of these devices indicate a businessman or woman or a student, but they have become so ubiquitous that it is difficult to assign such identities so easily. What kinds of things are in these various contraptions? Some are filled with their business documents and readings, especially since so many carry laptops so that they can fire up and work anywhere (in the year 2003, laptops began outselling personal computers) [14]. But they also carry many parts of personal archives. Photographs, credit cards, membership identifications, drivers’ licenses, and other items either include or represent the trail of documentary evidence that follows us everywhere we go, work, and play. With the laptops and digital cameras, most of us seem able to take nearly our entire personal archives with us wherever we travel, provided we have taken the time to scan in the older papers and photographs. It represents just one reason why digital curation might become as much a public concern. While efforts such as the PARADIGM project at the Universities of Oxford and Manchester are working on digital personal archives, recognizing that “there is a marked lack of research and development dedicated to the preservation of this kind of content,” since “to date, digital preservation projects have tended to be sponsored by corporate bodies or state archives.” Even so, this project has as its “principal audience . . . organizations, of any flavor, which care for the personal archives of politicians, scientists, writers, journalists, academics or of other individuals” [15]. My point is that we need to focus as well on the digital curation needs of average citizens, even if very little of their personal archives might ever reside in a repository.

Much of our personal space is occupied by older devices, from furniture to built-in shelving and filing cabinets, reflecting very traditional document systems. We reside in a world where we are bombarded by advertisements suggesting that most of this is obsolete or that it should be only of interest to us if we are engaged by the use of antiques and more comfortable with obsolete technologies (and I write this on a laptop on a hundred year old desk where sits a fountain pen and an old letter sorter). The maintenance of personal archives will be contested in this environment, raising many of the same issues archivists and journalists, academics or of other individuals” [15]. My point is that we need to focus as well on the digital curation needs of average citizens, even if very little of their personal archives might ever reside in a repository.

Much of our personal space is occupied by older devices, from furniture to built-in shelving and filing cabinets, reflecting very traditional document systems. We reside in a world where we are bombarded by advertisements suggesting that most of this is obsolete or that it should be only of interest to us if we are engaged by the use of antiques and more comfortable with obsolete technologies (and I write this on a laptop on a hundred year old desk where sits a fountain pen and an old letter sorter). The maintenance of personal archives will be contested in this environment, raising many of the same issues archivists and records managers face in striving to administer both paper and digital systems in our transitional era from analog to digital formats. How we engage such challenges and incongruities with our personal and family records ought to make us better in how we explain, engage with, and resolve such situations in other organizations supporting the archival programs we normally think of when we imagine the archival mission.

8. DIGITAL CHALLENGES

It is difficult to know just what the impact of the Internet (and other digital documentary forms) has been on the sense of the personal archive. As I have discussed in this paper, the Internet has provided a new space for displaying personal archives, such as blogs and personal photograph galleries, but there has been enough debate about the preservation of digital records that no one should view the newly emerging digital personal archives with unbridled confidence. Electronic mail has proved to be a particularly testy challenge, as it generates great gobs of
documents testing maintenance and control systems, posing new security problems, and, like the telephone before it, has probably lessened the reliance on the traditional letter. A child writing from college is now likely to ask for money by dispatching quickly an email, perhaps with digital images attached, a scan of a professor’s positive comments on an essay, and other documents. Saving these kinds of records suggests both greater problems and more imagination. No longer can we casually jam stuff into a box in the closet, planning to look at and assess the records at a later time. Now we must more consciously plan out what our personal archives will look like and the functions it will serve, understanding that it will consist of paper records, printed ephemera, photographs, memorabilia, and digital materials.

It is likely that the increasing use of digital formats will enhance interest in the preservation of personal archives and that this will strengthen the public’s awareness of the importance of archives, records, and information management. For example, as digital photography has captured the public’s interest, now constituting a billion dollar a year industry, with its ease of making and reproducing images, other problems have emerged. These images are easily manipulated, but just as easily lost. If archivists can step forward and provide advice about such problems, a new public comprehension of the archival mission may be gained.

Personal records and artifacts immerse us into the world in a way that gives us meaning beyond the superficial material stuff we acquire. All we have to do is to determine how best to care for the archival goods, recognizing that the personal sensibilities associated with family archives may be far more important than all the scholarship and theorizing being devoted to studying such documents. All we may need to do is to read the testimony of a writer like Ivan Doig rediscovering his mother and his childhood after he receives a packet of his mother’s letters years after her death, suddenly “sensing the carrying power of ink as a way to go on” [16]. We have to acquire a sense of the digital version of this “carrying power.”

9. WE ARE ALL ARCHIVISTS

Most people have little sense of what an archivist does, or even know that such a profession exists. It is difficult for the public to understand archival work because of the shifting notions of archives in scholarly and popular usage. A lot of scholarship about the nature of the "archive" has twisted and expanded the concept beyond any sense of the traditional meaning of the term as a repository where scholars and other researchers use historical records to try to decipher the past.

Fortunately, there are places to turn for assistance. Enough people struggle with administering their own personal papers that a number of self-help books have appeared offering advice and promising solutions. Even the vendors of the latest glitzy software products have gone after the market for managing personal records.

Maintaining one’s own personal or family papers is a crucial activity for preparing a family history, and while there are similarities between archiving and writing there are also differences. Americans have been blessed with an abundance of historical agencies providing many programs offering advice and instruction on family history work. And these organizations have been doing this for a long time.

Archivists can gain valuable assistance by understanding the nature of personal recordkeeping. For the archivist, some of these personal archives may ultimately be offered to their repositories, and it behooves the archivist to provide advice about the care of such documentary sources to ensure that they arrive in good order. Some professionals have awakened to this role, as reflected in the recent book by Don Williams and Louisa Jaggar on Saving Stuff, offering basic advice on the care of personal artifacts and documents [17]. Of course, the vast majority of personal records will not be offered to archives, as they are of much greater value to the individuals and families they relate to. But for the archivist, the key importance may be tapping into the concept of personal archives as a way of explaining the importance of administering records. Too often we assume that others do not share any sense of this important work, when, in fact, many are quite engaged in trying to preserve their own family and personal papers.

The archival profession needs to concentrate on developing new mechanisms for educating the public about how to care for their personal and family archives. Perhaps, this has already started to happen, as we note that one of the expanding sections of the major chain bookstores is the self-help section. In this section there is a growing number of books about organizing personal papers, what to keep, how to avoid cluttering up one’s life, preserving family archives, and so forth. Any of these self-help guides provides as good a reason as any why there is a need for people to follow the advice being offered. However, professional archivists can do better than most of the advice currently being offered in such publications, but, in order to do so, archivists need to shift some of their attention from mostly serving the needs of academic and experienced researchers to working with amateurs committed to preserving their personal and family archives.

Archivists need to be careful in how they might criticize the role, old or new, of individual preservers of documentary materials. A historical perspective clearly indicates that the source of many holdings in now established archival repositories is that of the work of individual collectors or the efforts of some family member to preserve the family archival legacy. While archivists often see or portray themselves as the documenters or collectors of our society, if the truth be told many of their holdings were already somewhat formed by individual collectors who built aggregations of documents or who worked to preserve their own family archives; their disposition in an established archival repository with a more public mission was simply the last stage of a process. What I am getting at is that we may be seeing a very different role for family archivists because of the digital platform on which they work. Whether this new archival desire emanates from simply a utilitarian interest in maintaining personal and family papers or whether it reflects a new kind of competition with archival repositories, it is too early to tell.

The transformation of archival perspectives has usually occurred because of rapidly growing volumes of documents and increasingly complex hybrid documentary technologies. Over the past two decades especially, new networked digital technologies have pushed archivists, at least some, to rethink the custodial model and to consider new kinds of distributed or post-custodial strategies. This seems to be where we now are with personal and family archives, and the prospects for continuing archival work are both daunting and exciting. Continuing archival work is
challenging because the digital recordkeeping and information technologies continue to perplex archivists, especially those working in smaller institutions with limited resources, in their abilities to apply traditional notions of record reliability and trustworthiness to the new environments; the result has been almost an avoidance of dealing with these documentary forms, an approach bound to cause problems with the future prospects of the archival profession. Dire predictions of the demise of the archival community have not come true, although it is also the case as much because there are still vast reservoirs of paper records to be analyzed and administered. It is unlikely that sanguine predictions will continue if archivists working through society simply do not deal with digital formats; this avoidance will result in other disciplines or new disciplines stepping in to fill the void.

The exciting aspect of rethinking how archivists will work in preserving personal and family archives is that it may re-open a much greater possibility for reaching the public with a clearer sense of the archival mission, an objective archivists and their professional associations have struggled to do for several decades with very mixed results. It is, however, clearly the case that the public itself is actually sowing the ground for archivists to seed. As individuals and families continually invest in new technologies that are portable and use them to store ever-growing amounts of records and information, they will encounter increasing challenges for maintaining these sources. Even as people grow more aware of the potential loss of these materials because of technical glitches and design weaknesses, they may be loath to give up on them because of their convenience and ease of use (similar to people being hesitant to stop using credit cards and retail discount cards even as they become more aware of the increased threats to personal privacy).

In addition to archivists adopting a broader campaign to assist the public, they must redirect part of their attention and resources from acquisition of archival materials and to assisting onsite researchers to developing workshops, self-help publications, and other tools for the purpose of equipping more and more citizens to care for their own archives. There is an emerging scholarly discipline, personal information management, archivists need to begin both to dig into and to influence (and there is probably a natural connection to the other emerging area, digital curation).

Partnerships with other disciplines to develop solutions for personal archives management and a greater dedication to research about the reasons why personal and family archives are formed and maintained are good commitments for the archival community to make. Archivists, either on their own or in collaboration with others, need to write and publish guides about the management of personal and family archives directed at the lay audience. Maybe we are becoming poised to do this as we begin to see new attempts to provide such publications. The Council of State Archivists, for example, published in late 2007 a guide on saving family records in disasters [18]. Advisory publications such as this are a by-product of what might be a new role for professional archivists. In this role archivists will function more as advisors rather than acquirers, educators giving their knowledge away rather than protecting the secrets of a guild, and advocates rather than reactors in seeking to preserve the portion of the documentary universe that possesses archival value. Some of this should seem familiar because it relates to some notions promulgated by those working with electronic records a decade or more ago, although some of these views have been greeted with criticism or silence.

I am not arguing that established, institutional archives will not acquire and preserve personal and family archives in the future. What I am suggesting is that the vast and rapidly growing digital documentary universe of such archives requires that archivists first try to advise the creators and amateur caretakers of these materials and only intercede when valuable or unique personal and family archives are endangered. This will require new and more intense archival appraisal approaches, ones that have not been devised yet, as well as new standards for the maintenance and use of personal and family digital archives. These are interesting and engaging problems that should be the increasing focus of what educators teach, what doctoral students conduct research about, and what working archivists experiment with and develop into reliable systems.

10. WHAT DIGITAL CURATION?

The digital age is such that we read of new terms, concepts, jobs, and even disciplines on a regular, sometimes frightening basis. Digital curation is, of course, just another new entry in our new cyberuniverse, and the question of whether it has staying power or not depends on how relevant and practical it proves to be in supporting archivists and their allies to develop new training programs both for records professionals and for private citizens interested in preserving digital documentation.

The efforts in the digital curation project to catalog required knowledge and to define and explain this knowledge with specificity in order to build a logical curriculum are quite important. Terms like “systems engineering and development,” “harvesting,” “ingest,” “digital objects or packages,” and so forth, all seem perfectly at home in a graduate-level curriculum. However, there also needs to be some thought given as to how to condense or spin-off elements of this curriculum, minus the technical or professional jargon, in order to create guides, training programs, and other forms of advice for public use and consumption.

My hope is that we can build on the digital curation initiative in order to develop the means by which to help private citizens to care for their materials. However, even thinking in this way might be far too limited. Nicholas Carr, in his latest book, predicts the rise of private computer systems companies providing centralized data processing services and storage. Carr also hints that these services will not just be directed at large companies with their immense computing needs, but that individuals might also be in line to use such services. Already, Carr suggests, “our PC’s are turning into terminals that draw most of their power and usefulness not from what’s inside them but from the network they’re hooked up to – and, in particular, from the other computers that are hooked up to that network” [19]. Some archivists might think that this means they should contemplate developing such a service role, but I believe that would involve resources far beyond the archival profession. However, it might mean that archivists should target these emerging computer services and data companies to educate them about the archival issues. The digital curation definitions and curriculum might provide one means to do this.
ACKNOWLEDGMENTS
This is a shortened version of a book by the author, Personal Archives and A New Archival Calling: Readings, Reflections and Ruminations, recently published by Litwin Books, LLC.

11. REFERENCES
Invited Paper: Use of Computer Forensics in the Digital Curation of Removable Media

William Bradley Glisson
Humanities Advanced Technology and Information Instituted (HATII),
The University Glasgow, Scotland
b.glisson@hatii.arts.gla.ac.uk

ABSTRACT
The purpose of this paper is to encourage the discussion of the potential place and value of digital forensics techniques when dealing with acquisitions on removable media in the field of digital curation. It examines a basic computer forensics process, discusses a typical file system for removable media, and raises questions about necessary processes and incentives for addressing data capture in the field of digital curation.

Keywords
Computer, computer forensics, digital, digital forensics, digital curation, archival

1. INTRODUCTION
As has been noted in past papers, information is an extremely valuable asset in the world's increasingly globally networked environment [1]. The digital revolution has contributed to a radical paradigm transformation in today's information rich societies. Many of society's basic operational components such as financial markets, health care, and governmental agencies depend on information in digital formats.

An excellent example of the growth in digital information is provided by the White House. In 1978, Congress passed the Presidential Records Act, “which requires each president to maintain records of all activities, deliberations, decisions and policies that reflect on performance in office” [2]. There have been questions in the press as to whether the National Archives can handle the volume of digital data produced during the Bush administration [3]. It has been estimated by archive officials that the “… electronic record(s) of the Bush years (are) about 50 times as large as that left by the Clinton White House in 2001” [3].

This paper discusses the potential place and value of digital forensics techniques for collecting institutions (e.g., libraries, museums, archives) that are dealing with acquisitions on removable media. The hope is that this paper will encourage future research in the adoption of computer forensics processes and tools for digital curation.

2. COMPUTER FORENSICS PROCESS
This section summarizes a typical digital forensics workflow.

2.1 Process for Retrieving Data
For this discussion, consider the forensic acquisition of a data key, a.k.a. thumb drive. One must first decide where to store the information. In order to counter data remembrance so that it does not contaminate the information stored on the target drive, the target drive needs to be forensically cleaned. This means that the target drive is wiped by writing all zeros or ones to the drive. Many companies promote a US Department of Defense (DOD) standard on this topic. However, the 2006 National Industry Security Program Operating Manual (that is also referenced as the DOD 5220.22-M) does not specify the number of passes required to achieve sanitation [4].

The Defense Security Service (DSS) Clearing and Sanitization Matrix which was updated on June 28, 2007, makes the statement that “DSS will no longer approve overwriting procedures for the sanitization or downgrading (e.g. release to lower level classified information controls) of IS storage devices (e.g., hard drives) used for classified processing” [5]. There is clearly controversy over the effectiveness of overwriting drives. A recent study that claims “…that correctly wiped data cannot reasonably be retrieved even if it is of a small size or found only over small parts of the hard drive” [6]. Even thought there appears to be a doubt as to the effectiveness of overwriting for sanitation purposes, it is still a good idea from a forensic practice perspective.

The next activity is to document the hardware. This includes any serial numbers and manufacturer information. The next activity in a forensic situation would be to start the chain of custody and to transport the device to a secure lab for processing.

At this point, a bit-stream copy of the removable media should be made by creating either a clone or a forensic image of the device. A bit-stream copy of the removable media copies every bit on the source drive [7]. When a bit-stream copy is saved to another drive, i.e., the target drive (generally, this drive needs to be identical to the source) so that the target drive is bootable, this is commonly referred to as a clone. When the bit-stream copy is saved to an image file, this is commonly referred to as a forensic image. It is possible to take a forensic image and restore the image to a drive making a clone of the source drive.

At this point, the forensic copy of the removable media then needs to be authenticated. This is typically done through the execution of a one-way hash on both devices to verify that they are identical.

The next step is the analysis of the drive to identify active files and inactive files. Active files are readily identifiable and can be accessed with the appropriate software and, in some cases, the required security information. Inactive files can be located by carving the unallocated space and slack space off of the drive. Unallocated space is space that has not been used by the file system. It can contain deleted documents in Windows and DOS
operating systems. Information can also be found in two types of unallocated slack space: file slack and RAM slack (sometimes both are referred to as drive slack) [7]. Any anomalies that are identified such as encrypted information, proprietary software formats and missing partitions are noted and examined individually. All of the information that is found would be documented appropriately.

The documentation is very detailed so that it includes all of the issues that were encountered and the evidence that was discovered in the process. It will also include the methods utilized in the investigation along with citations supporting the analysts’ stated opinions.

2.2 Removable Media File Systems
The next issue to address is the file system. It can be argued that the file system is part of the application layer, the presentation layer and the session layer as defined in the Open Systems Interconnection (OSI) seven-layer model [8]. The file system is responsible for the organization of the files, i.e., it is responsible for the logical placement of the files on the storage drive. Hence, the file system is manipulating the sectors on a drive so that they are treated as clusters. These clusters are then linked together, as needed, so that they can be treated as a file with associated metadata. The size of the clusters will vary depending on the size of the hard disk and the file system [7]. Understanding this interaction is critical to the retrieval of data that has been accidentally or intentionally deleted on various types of files systems like the File Allocation Table (FAT) system, New Technology File System (NTFS), High Performance File System (HPFS) or the Hierarchical File System (HFS).

It is common, although not mandatory, for data keys to use a version of the File Allocation Table (FAT) system. When a file is deleted in a FAT system, the first character of the file is replaced with a non-readable character and the FAT entries linking the sector clusters are zeroed out. The data still exist on the system. The zeroing out of the entries linking the clusters simply tells the file system that the space is available for use. Hence, to restore a file, the name of the file would need to be amended and the links between the sector clusters would need to be re-established. If additional data has been saved to the system after a file has been deleted, the old data may have been over written.

3. INFORMATION APPLICATION
Now that we have an understanding of the basics of computer forensics, how can we apply this to the field of digital curation? In the case of the White house and the Presidential Records Act, does the administration have an obligation to not delete information? Does the archivist have an obligation to recover as much information as is possible from the digital media provided by the administration? If so, how does the archivist achieve this goal?

4. CONCLUSION & FUTURE WORK
This paper is intended to raise awareness of computer forensics concepts and to prompt discussion about the potential use and the need for computer forensics processes and tools in the field of digital curation. The long-term implications, obstacles, and hurdles for the integration of digital forensics processes and techniques into the field of digital curation are not fully understood. It is clear that the digital revolution will continue to penetrate all aspects of our globally networked information rich society. This continued integration raises the need to address the amount of information that is archived along with the examination of the processes that are implemented.

There are many associated questions to consider:

- When collecting institutions receive removable media, what procedures do they follow?
- Do they only capture “live files,” or do they also capture deleted files?
- Under what conditions would it be beneficial for collecting institutions to copy entire drives, i.e. all of the bits, rather than only copying the live files from the drives?
- Are current practices of collecting institutions practical from a business perspective? To what extent do they conform to established digital forensics principles and practices?
- How often, and under what conditions, will the processes and storage arrangements of collecting institutions need to be upheld in a court of law?

Future research could focus on a closer examination of the process used in the field of digital curation. This can include conducting survey inquiries with collecting institutions in several different countries. It could also include the investigation and development of a digital recovery methodology specifically for use in digital curation. It could reasonably include a targeted educational effort toward librarians and archivists on relevant tools and the operation of specific file systems. The educational effort could be followed by empirical studies of the practicality and effectiveness of the developed methodologies to meet the needs of collecting institutions and their target user communities.

5. REFERENCES
Moving Web Archiving into the Classroom

Molly Bragg
Internet Archive
San Francisco CA, 94129
USA
mbragg@archive.org

Edward A. Fox
Computer Science
Virginia Tech
Blacksburg, VA 24061 USA
fox@vt.edu

Margaret Hedstrom
School of Information
University of Michigan
Ann Arbor, MI 48109 USA
hedstrom@umich.edu

Christopher A. Lee
School of Information & Library Science
University of North Carolina
Chapel Hill, NC 27599 USA
callee@ils.unc.edu

ABSTRACT
Web archiving is a professional activity that manages the risk of information loss by identifying, selecting, collecting, managing and providing ongoing access to resources from the Web that are deemed to have sufficient continuing value. Graduate-level professional education has only recently begun to address the role of web archiving within digital collection and digital library development. In order to promote and cultivate such “citizen web archiving,” it will be beneficial for information professionals to provide opportunities for training and guided hands-on experience with available tools. The members of this session panel have led efforts to introduce web archiving into university and K-12 classrooms. Based on their experiences from the educational activities discussed above, members of the panel will discuss the following questions: What are the main principles, concepts and skills required to appropriately archive web resources? What are the most effective strategies for and biggest challenges associated with incorporating web archiving activities into the classroom?

Categories and Subject Descriptors

Keywords
Web archiving, professional education, collection development, web crawling, digital curation.

1. INTRODUCTION
All sectors of society are generating, disseminating and making extensive use of Web resources, but perpetual access to many of those resources is at risk. Web archiving is a professional activity that manages the risk of information loss by identifying, selecting, collecting, managing and providing ongoing access to resources from the Web that are deemed to have sufficient continuing value [1]. Rather than targeting individual files for capture, web archiving is usually carried out with software that either submits queries to designated sites or recursively follows links from designated “seed” Uniform Resource Locators (URLs).

Several large-scale initiatives – including the Internet Archive, Preserving and Accessing Networked Documentary Resources of Australia (PANDORA), Stanford WebBase, and Kulturarw3 of the Swedish Royal Library – have been archiving web sites for more than a decade. There is now a substantial body of expertise related to web archiving, but most of that expertise resides within a handful of large institutions.

There are now many available tools and services – both free and fee-based – that are designed to support web archiving. These include crawlers, indexing tools, search utilities, as well as entire software suites that allow users to specify, schedule, execute and monitor web-collecting activities.

In recent years, numerous collecting institutions have begun to investigate and implement strategies for archiving web resources that fall within their collection missions. There appears to be an increasing awareness of available tools and services, but very few information professionals have been trained to select, configure, apply or evaluate the tools and services. Graduate-level professional education has only recently begun to address the role of web archiving within digital collection and digital library development.

Outside the context of professional curation, private individuals have the potential to contribute substantially to social memory, by informing, contributing to, and initiating web archiving activities. Many bookmarking services and browser plug-ins allow individual users to identify specific web resources to be archived. Several Web archiving projects – from the 2-billion-page crawl of the Internet Archive in 2007, to the very focused collecting initiatives of many other organizations based on specific topics or events – have also called on individuals to suggest resources to be included within web archive collections. In order to promote and cultivate such “citizen web archiving,” it will be beneficial for information professionals to provide opportunities for training and guided hands-on experience with available tools.
2. MOVING WEB ARCHIVING INTO THE CLASSROOM
The members of this session panel have led efforts to introduce web archiving into the classroom.

In 2007, the School of Information at the University of Michigan introduced a class called Web Archiving (SI 639), which exposes students to tools and appropriate techniques for preservation of information delivered through both the “surface” and “deep” Web, including both capture and transformation into persistent formats and data structures.

Since 2006, the School of Information and Library Science at the University of North Carolina at Chapel Hill has been providing students with hands-on experience in web archiving, through assignments in classes on electronic records management (INLS 525) and collection development (INLS 513).

At Virginia Tech, web archiving efforts were launched after the shooting on April 16, 2007 [2]. Students have continued work in connection with an NSF-funded research study (http://www.dl-vt-416.org/). Early in 2008, this work was extended to include coverage related to the shooting at Northern Illinois University. In Fall 2008, the Department of Computer Science has been incorporating web crawling tasks and study of the related literature were integrated into a course on Digital Libraries (CS 6604). This is part of an NSF-funded partnership between Virginia Tech and the University of North Carolina (led by Jeffrey Pomerantz and Barbara Wildemuth) to develop a digital library curriculum (http://curric.dlib.vt.edu/). Students are helping by developing one module on “File formats, transformation, migration” and another on “Web archiving.”

In the spring of 2008, the Internet Archive, Library of Congress and California Digital Library initiated the K-12 Web Archiving pilot program, which allowed students from three different high schools in different parts of the country to select Web content to be included in “time capsules” that document aspects of contemporary life from a student’s perspective. In October 2008, the project is expanding to include 10 participating schools, including several high schools (including one Spanish-speaking school), as well as middle and elementary schools. The students are using the Internet Archive’s Archive-It web archiving service to create the time-capsule collections.

3. QUESTIONS FOR DISCUSSION
Based on their experiences from the educational activities discussed above, members of the panel will discuss the following questions, and others, as well as address concerns and issues raised by the audience: What are the main principles, concepts and skills required to appropriately archive web resources? What are the most effective strategies for and biggest challenges associated with incorporating web archiving activities into the classroom?

4. REFERENCES
Invited Panel: Gaps and Persistent Challenges

Clifford Lynch
Coalition for Networked Information
Washington DC, 20036 USA
clifford@cni.org

Kevin Ashley
Digital Archives Department
Univ. of London Computer Centre
London, WC1N 1DZ UK
K.Ashley@ulcc.ac.uk

Donald Sawyer
VIE, Inc.
Greenbelt MD 20771 USA
sawyer@acm.org

Raymond J. van Diessen
IBM Business Consulting Services
1066 VH, Amsterdam, The Netherlands
raymond_vandiessen@nl.ibm.com

ABSTRACT
Where have we been? Where do we need to go to advance the principles and goals of digital curation? This invited panel brings together leading researchers and thinkers in the areas of digital curation and digital preservation to respond to this critical question. Panelists will offer their unique and informed perspectives on the scope, extent, relevance, and quality of current digital curation research, practices and literature.

Keywords
Digital curation, digital preservation, data curation.

This work is licensed under the Creative Commons Attribution-Noncommercial-No Derivative Works 3.0 Unported license. You are free to share this work (copy, distribute and transmit) under the following conditions: attribution, non-commercial, and no derivative works. To view a copy of this license, visit http://creativecommons.org/licenses/by-nc-nd/3.0/.

DigCCurr2009, April 1-3, 2009, Chapel Hill, NC, USA
In Invited Panel: Personal Digital Archiving

Jeremy Leighton John  
British Library  
London, NW1 2DB UK  
Jeremy.John@bl.uk

Catherine C. Marshall  
Microsoft Research  
Mountain View, CA 94943 USA  
cathymar@microsoft.com

David Pearson  
Digital Preservation  
National Library of Australia  
Parkes Place, ACT 2600 Australia  
dapearso@nla.gov.au

Andreas Rauber  
Vienna University of Technology  
A-1040 Vienna, Austria  
rauber@ifs.tuwien.ac.at

ABSTRACT
This panel addresses personal digital archiving, which as an essential, but often neglected, arena of digital curation. Panelists will discuss their efforts to understand, support and document personal digital archiving activities, as well as exploring general challenges and opportunities for digital curation professionals.

Keywords  
Personal digital archives, personal information management, personal collections, digital curation.

1. INTRODUCTION
Individuals as well as small businesses (small office/home office, SOHO) have tremendous amounts of digital information. At the same time, they have little to no expertise on how to manage it, not to mention caring for their long-term preservation, as even simple back-up strategies pose already drastic challenges.

We thus need new approaches to fully automated digital preservation systems that are suited to their needs. Very much like a software firewall and antivirus software packages cannot replace an IT security department, but still fulfill the needs of many users, a simple software solution automatically managing digital preservation tasks may offer a feasible solution to a drastic upcoming problem for society at large.

2. DIGITAL LIVES AND E-MANUSCRIPTS
The Digital Lives Research Project is designed to provide a major pathfinding study of personal digital collections [1]. The project team drawn from the British Library, University College London and University of Bristol is led by Dr Jeremy Leighton John of the British Library (the lead partner) and funded by the Arts and Humanities Research Council (AHRC).

The Digital Lives team expect outcomes from their research to generate significant interest within arts and humanities, and the archives, library and information science professions as well as social and human sciences. It will also be of direct relevance to individuals who wish to manage their own personal digital collections for family history, biographical or other purposes.

This work is licensed under the Creative Commons Attribution-Noncommercial-No Derivative Works 3.0 Unported license. You are free to share this work (copy, distribute and transmit) under the following conditions: attribution, non-commercial, and no derivative works. To view a copy of this license, visit http://creativecommons.org/licenses/by-nc-nd/3.0/.

Jeremy Leighton John will share lessons and observations from the Digital Lives project, including First Digital Lives Research Conference, which was held at the British Library on February 9-11, 2009. He will also share experiences from the British Library’s acquisition and management of born-digital manuscript collections.

3. PERSONAL DIGITAL ARCHIVING TECHNOLOGY PRACTICES AND NEEDS BASED ON EMPIRICAL DATA
Catherine Marshall has conducted numerous studies of personal information management and personal archiving [2][3]. Early personal digital archiving research has been predicated on three assumptions, all of which are worth a second look as we gather more experience living in a digital era. The first is that our digital belongings are by-and-large a collection of local files; the second is that personal archives will be centralized repositories; and the third is that digital stewardship is a matter of scaling down institutional best practices to suit the needs of the individual. While there is a grain of truth to each of these assumptions, it is dangerous to embrace them without question. Marshall will present the empirical evidence that casts doubt on the assumptions, and will explore briefly what the implications of this evidence might be for personal digital archiving technologies.

4. A CALL FOR COLLABORATIVE DEVELOPMENT
David Pearson has had practical experience at the National Library of Australia and National Archives of Australia in recovering data from hundreds of obsolete as well as contemporary media carriers in both Government and Personal Archives [4]. This experience has provided Pearson with an understanding of the issues which can be encountered, including assessing which materials can be processed, understanding significant properties and finding suitable renderers for file formats, recovering data from degrading carriers, and locating suitable hardware and software to read the carriers. These issues already make it difficult for digital preservation solutions to preserve access to digital content.

Pearson has gained significant experience and knowledge from his involvement in a number of digital preservation projects that seek to address these issues. Although most organizations have slightly different business drivers, Pearson and his colleagues believe that from a preservation and risk management perspective, we must approach the problem (of preserving digital material) as a community, or we ignore it at our individual peril.
The requirements in data management as well as managing the preservation task fully automatically pose significant research challenges. Solving these will represent an important step forward in enabling non-experts to preserve their digital data in very much the same way and at acceptable levels of quality as they can currently preserve conventional objects for their own needs.

5. HOME OFFICE PAINLESS PERSISTENT LONG-TERM ARCHIVING (HOPPLA)
Andreas Rauber will discuss HOPPLA (Home and Office Painless Persistent Long-term Archiving), which combines back-up and fully automated migration services for personal data collections and SOHO settings [5]. It will combine bit-stream preservation via LOCKSS-style back-up with logical preservation by automatically obtaining migration rules and tools. These tools will be provided based on the local collection profile via an update service similar to and inspired by current anti-virus software solutions. Data is ingested from a number of sources such as data carriers, email repositories and on-line storage locations such as wikis and blogs, while back-up storage supports off-line and on-line storage media in both write-once as well as rewritable forms.

6. DISCUSSION AND IMPLICATIONS
The panelists will discuss the implications of their various experiences and research findings for the practice of personal digital archive curation. Issues include design requirements for software and systems; needs or professional development; and realistic goals for advancing the interests of individuals now and in the future.

7. REFERENCES
ABSTRACT

This panel’s speakers will discuss policy issues associated with the curation of digital materials, including privacy, copyright/intellectual property, cultural sensitivity and trustworthiness. Several collection types will be included in the discussion including health information, multimedia collections, scientific data collections and digital library, museum and archival cultural collections.

Categories and Subject Descriptors

K.4.1 [Public policy issues]: Intellectual property rights, privacy, ethics, regulation; K.5.1 [Hardware/Software protection]: Proprietary rights; K.5.2 [Governmental issues]: Regulation

Keywords: HIPAA, privacy, copyright, data sharing, intellectual property, data collections, multimedia collections, health information, rights management; deidentification; digital rights management; DRM

1. INTRODUCTION

Curators of digital cultural collections confront a variety of policy issues including questions of privacy and personally identifiable information, copyright and intellectual property concerns, sharing of culturally sensitive materials, and the trustworthiness of collaboratively developed and managed collections. This panel will feature five speakers whose work has explored policy aspects of curation of digital materials – both born digital materials and digitized collections of analog materials. Catherine Arnott Smith will discuss the impact of HIPAA on the management and digitization of historical medical collections. Grace Agnew will describe rights issues associated with multimedia and performance collections. Melissa Cragin will describe the preliminary results of her investigations of curation issues associated with scientific data collections – including intellectual property issues and issues of trust. Jean Dryden will discuss the technical and non-technical measures that Canadian archival repositories use to control further uses of their archival holdings digitized and made available online (even though in many cases they aren't the copyright owners). Kristin Eschenfelder will describe the results of an IMLS funded study of U.S. cultural institutions use of rights management technologies to control access and use of collections for copyright, privacy or other reasons.

2. Health Information Collections and HIPAA [Arnott Smith]

A truism of medical information is that it is only useful when it is shared. Confidentiality, privacy, and security, as legislated through HIPAA [1], pose significant challenges not only to management of contemporary, but historic health information, since HIPAA does not exempt old material from regulation; it is the first federal legislation that protects the dead. One common type of health information record is the patient or medical record. Certain content in the medical record may be potentially embarrassing or socially stigmatizing. Both active and historic records generated in the process of healthcare delivery are potentially controversial and regulated by law. A further complication is the fact that although medical records exist to document the encounter between a single patient and a healthcare system, more individuals than simply the patient are legitimately represented in the record. People exist embedded in genetic and socially constructed networks impacting their healthcare status and their ways of life. For example, patients who live with other people – a spouse, a partner, a grandparent, a child—share living conditions, which may be important to the continued care of the patient. It is nontrivial, for example, if an asthmatic woman has a spouse who is unable to quit smoking. The asthmatic woman’s spouse will almost certainly be documented in “her” medical record.

This work is licensed under the Creative Commons Attribution-Noncommercial-No Derivative Works 3.0 Unported license. You are free to share this work (copy, distribute and transmit) under the following conditions: attribution, non-commercial, and no derivative works. To view a copy of this license, visit http://creativecommons.org/licenses/by-nc-nd/3.0/.

DigCCurr2009, April 1-3, 2009, Chapel Hill, NC, USA

CC BY NC ND

117
For both stigma and social-network reasons, then, there are considerable social implications in the management of health information. What are some inherent problems in digital curation of this socially sensitive medical data—old and new? In this paper, I present examples of problematic materials in medical archives and libraries, as well as results from a study using Encoded Archival Description and documents’ internal knowledge structures to enhance use of these problematic materials.

3. Policy Issues Association and Video Collections [Agnew]

Curation of multimedia and performance collections present difficult intellectual property issues because of the mix of still and moving visual images and audio works that make up audiovisual materials, and also because of the complex layers of rights associated with performance works often included in multimedia collections. Agnew, author of the recent book, Digital Rights Management: A Librarian's Guide to Technology and Practice will reflect on the policy issues she experienced during her extensive experience with the development of multimedia collections including the Moving Image Collections (MIC) portal, and with NJVid, the digital video portal for the state of New Jersey. Agnew will describe policy, management and technology tools available to assist cultural institutions in their efforts to proactively tackle rights issues associated with multimedia and performance collections.

4. Scientific Data Collections [Cragin]

Scientific data collections (SDCs) present digital curators with issues of data management, scientific communication and collective activity. They are more complex than “just” databases, being themselves simultaneously the subjects and objects of research, and can illuminate the work processes involved in knowledge production. As such, they offer a window on the convergence of emergent forms of knowledge production; negotiation of organizational structures; technical infrastructures; and the interactions of the multidisciplinary needs and the expertise of users. This is a rich domain for studying issues inherent in knowledge production by human beings: intellectual property concerns and questions of quality and trust. For example, how do professionally conflicting notions of confidentiality, data ownership, and data sharing resolve when the data resources under discussion are public data resources? How do differences in understanding of what constitutes data “quality” affect development of metadata? And the resolution of these issues has implications for handling of SDCs by digital curators who must appraise and select these complex collections for their archives and libraries.

5. Copyright and Archival Collections [Dryden]

Dryden will discuss an aspect of her doctoral dissertation, which investigated the copyright practices of Canadian archival repositories in digitizing selected documents from their holdings and making them available on the Internet [3]. Her dissertation findings suggest that Canadian archival repositories are invoking copyright in ways that may impede or discourage access to, and use of, their online holdings, even if they have no copyright interests to protect. Although repositories may, in certain situations, be authorized to administer copyright interests in their holdings, they generally have no legal or professional obligation to enforce others’ copyright interests. Nonetheless, the dissertation found that they attempt to control further uses of their online holdings through the use of technical measures (e.g., watermarks, low-resolution images, disabling the right-click), non-technical measures (e.g., statements specifying terms and conditions of use, indications of copyright status or ownership in structured descriptions, and copyright guidance for users), or both. While the research suggests that this may be due to factors other than copyright (e.g., a desire to generate revenue, professional obligation to convey the archival context of the document(s) in question, or to promote or safeguard the reputation of the repository), controls on further uses are often presented under the rubric of copyright. Access to online documentary heritage may be inappropriately limited as a result. Dryden’s presentation will discuss possible ways to mitigate these limitations, and fruitful avenues for further research.

6. Controlled Online Collections [Eschenfelder]

Eschenfelder reports on the results of an IMLS funded study of U.S. library, museum and archive controlled online collections (COC) [4]. A COC includes digital materials for which access is limited or use is limited. For example, a COC may limit access to only a certain subset of the public. Or, a COC may limit certain uses – for example disallowing local saving of digital materials. Cultural institution’s motivations for creating COC include (but are not limited to) concerns about copyrights or the difficulty of obtaining rights clearances, concerns about the privacy of individuals depicted in materials, the desire to limit the exposure of culturally exclusive materials such as photographs of religious ceremonies, the need to protect fragile historical or ecological sites, and revenue generation. Eschenfelder will report on the types of management, policy and technical tools employed by cultural institutions to control access to and use of collections, and the types of collections that cultural institutions seek to protect. Eschenfelder will also describe a number of innovative projects that employ rights management technologies to provide limited online access to materials that might not otherwise be available on the Internet.

7. REFERENCES

Invited Paper: “Able To Develop Much Larger and More Ambitious Projects”: An Exploration of Digital Projects Teams

Lynne Siemens  
Faculty of Business  
University of Victoria  
Victoria, British Columbia, Canada  
siemensl@uvic.ca

Wendy Duff  
Faculty of Information Studies  
University of Toronto  
Toronto, Ontario, Canada  
wendy.duff@utoronto.ca

Richard Cunningham  
Acadia Digital Culture Observatory  
Acadia University  
Wolfville, Nova Scotia, Canada  
richard.cunningham@acadiau.ca

Claire Warwick  
Department of Information Studies  
University College London  
London, UK  
c.warwick@ucl.ac.uk

ABSTRACT
This paper reports on a research project which is exploring the prevalence and nature of research teams undertaking digital projects. Drawing upon interview and survey data, it aims to identify the methods and patterns of interaction used by collaborating digital projects teams and provide recommendations to support effective and efficient teams.

Keywords  
Digital libraries, digital archives, teamwork, collaboration, project management, digital projects.

1. INTRODUCTION
This paper reports on a research project which is exploring the prevalence and nature of research teams undertaking digital projects. Drawing upon interview and survey data, it aims to identify the methods and patterns of interaction used by digital projects teams in their collaborations and to provide recommendations to support effective and efficient teams.

2. CONTEXT
Traditionally, research contributions in the scholarly field have been felt to be, and documented to be, predominantly solo efforts by academics involving little direct collaboration with others, a model often reinforced through doctoral studies and beyond [See, for example, 7, 19]. However, work within the Digital Humanities and Libraries communities is an exception to this. Given that the nature of project work involves computers and a variety of skills and expertise, members of these communities are working collaboratively within their institutions and with others nationally and internationally to undertake their work. Such collaboration typically involves the need to coordinate efforts between librarians, academics, undergraduate and graduate students, research assistants, computer programmers, content experts, and other individuals as well as the need to coordinate financial and other resources. However, too little is known about how these teams work or the supports they need to be successful.

That said, efforts toward understanding the organizational context in which digital projects are situated is beginning in earnest. Two large-scale survey projects [26, 30] have highlighted issues of collaboration, among other topics, and Warwick [32] found that the organizational context has had an impact on the manner in which Digital Humanities/Humanities Computing centres developed in the United States and England. An initial qualitative research project found that Digital Humanities research is accomplished within research teams, which are developing tools and processes to facilitate that collaboration [25]. In addition, McCarty [17] explores the ways that computers have opened the opportunities for collaboration within the humanities and has explored the associated challenges of collaboration and team research within the HUMANIST listserv [16]. Individuals in the field are also reflecting on their own experiences in teams through conference presentations and papers [See for example: 13, 14, 21-23, 27, 31]. Finally, through efforts such as the University of Victoria’s Digital Humanities Summer Institute, Irish Royal Academy’s Digital Humanities Observatory Summer School, the UNB/Acadia University Fall Institute for Digital Libraries and Humanities, and other similar ventures, the community is working to develop its collaborative capacity through workshops in topics like community-specific project management skills.

This study draws and builds upon these efforts as it explores and formally documents the nature of research teams within these communities to the end of identifying exemplary work patterns and larger models of research collaboration that have the potential to strengthen this positive aspect of the community even further.

3. METHODS
This research project used a two-pronged inductive approach with a combination of data collection methods. First, members of various multi-disciplinary, multi-location research teams in Canada, United States, and the United Kingdom were interviewed. Lasting about an hour, these interviews explored the individual’s research team context with a focus on the...
participants’ definition of teams; their experiences working in teams; and the types of supports and research preparation required to ensure effective and efficient research results. Second, drawing upon themes from these interviews, a survey of members of the general Humanities Computing community was undertaken in order to establish the prevalence of research teams within the community. It was distributed to members of the Society for Digital Humanities/Société pour l'étude des médias interactifs, Association for Literary and Linguistic Computing, Association for Computing and the Humanities, Centrenet and HUMANIST listserve. As time of writing, the survey is also being sent to members of the Digital Libraries community. The survey provides descriptive statistics on the number of teams, their composition, and perceived effectiveness; it also establishes a baseline against which further research in this field and others can be compared. The results include a description of the community’s work patterns and relationships and the identification of supports and research preparation required to sustain research teams [as per 15, 18]. This two-pronged approach corresponds to similar studies [1, 4, 8, 11].

As discussed below, the resulting sample is small and self-selected. The respondents are more likely to be involved in digital humanities teams. Further research will enlarge the size and scope of the sample.

4. RESULTS
At the time of writing this paper, final data analysis of the survey and interviews is being completed. Preliminary results can be reported at this time with final results being discussed at the conference.

The seven interviewed individuals have participated in a range of team research projects, in terms of project objective, team membership size, budget, and geographical dispersion, both within their own institution, nationally, and internationally. The roles they play are varied and include research assistant, researcher, computer programmer, project manager, and lead investigator. Despite the diversity among their teams and roles, these individuals share several commonalities in terms of their methods of interaction and collaboration within their research teams.

In total, 36 individuals responded to the survey. Like the interviewees, they have been involved in a variety of digital projects, ranging from digitized manuscripts, electronic editions, databases, software and others. Table 1 summarizes survey respondent demographics.

The survey results provide elaboration on the methods of interaction and collaboration within the digital projects community. The majority of the respondents have been involved in two or more digital projects of various types in the past year. Given the many challenges associated with team projects (as discussed below), individuals who undertake this work must derive substantial benefits. Within this survey, the top reasons for team work include:

- Team members have different skill sets (90%)
- Collaboration is more productive than individual work (63%)

<table>
<thead>
<tr>
<th>Demographic</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Affiliation (36)</td>
<td>University (32)</td>
</tr>
<tr>
<td></td>
<td>Research Centre (1)</td>
</tr>
<tr>
<td></td>
<td>Other (3)</td>
</tr>
<tr>
<td>Language (34)</td>
<td>English (28)</td>
</tr>
<tr>
<td></td>
<td>French (4)</td>
</tr>
<tr>
<td></td>
<td>Spanish (1)</td>
</tr>
<tr>
<td></td>
<td>Other (1)</td>
</tr>
<tr>
<td>Role within the Team (32)</td>
<td>Leader (22)</td>
</tr>
<tr>
<td></td>
<td>Team Member (6)</td>
</tr>
<tr>
<td></td>
<td>Technical Support (1)</td>
</tr>
<tr>
<td></td>
<td>Other (4)</td>
</tr>
<tr>
<td>Team’s Membership Size (33)</td>
<td>2 individuals (1)</td>
</tr>
<tr>
<td></td>
<td>3-5 individuals (15)</td>
</tr>
<tr>
<td></td>
<td>6-10 individuals (6)</td>
</tr>
<tr>
<td></td>
<td>More than 10 individuals (11)</td>
</tr>
<tr>
<td>Budget (32)</td>
<td>$0-9,999 CDN (6)</td>
</tr>
<tr>
<td></td>
<td>$10,000-49,999 CDN (5)</td>
</tr>
<tr>
<td></td>
<td>$50,000-249,999 CDN (9)</td>
</tr>
<tr>
<td></td>
<td>$250,000 CDN and more (12)</td>
</tr>
<tr>
<td>Position (22)</td>
<td>Academic (10)</td>
</tr>
<tr>
<td></td>
<td>Post Doctoral Fellow (2)</td>
</tr>
<tr>
<td></td>
<td>Programmer/Developer (4)</td>
</tr>
<tr>
<td></td>
<td>Researcher (2)</td>
</tr>
<tr>
<td></td>
<td>Student (4)</td>
</tr>
<tr>
<td>Academic Discipline (34)</td>
<td>Humanities (19)</td>
</tr>
<tr>
<td></td>
<td>Social Sciences (5)</td>
</tr>
<tr>
<td></td>
<td>Library and Information Sciences (4)</td>
</tr>
<tr>
<td></td>
<td>Computer Science (6)</td>
</tr>
</tbody>
</table>

- Different methodological approaches to research by team members (54%)
- The project schedule requires multiple staff to complete it on time (51%)
- The volume of data to be studied (42%)
- Enjoyment of collaboration (42%)

Additionally, one of the interviewees remarked that grant funding agencies were also playing a role, by demonstrating a preference towards projects that require a team approach to complete.

These reasons were further emphasized in open-ended questions regarding the benefits of team work. The general themes of these answers focused on the development of a team spirit, the ability to be more productive and efficient and undertake more complex and larger projects, the ability to combine different expertises, skill sets, and perspectives within the project, and the opportunity to learn from collaborators. The interviewed individuals also echoed these comments, with a particular emphasis on the social and fun aspect of team work. In addition, one interviewee also suggested that an additional benefit of digital team projects is the diffused sense of responsibility with a team. Given that the work is undertaken by a group of individuals, not all the work and outcomes are “on one’s shoulders if it flops.” This interviewee felt that this context provides an opportunity for a team to be more experimental within the project. Given these various benefits, some interviewees acknowledged that they undertook most, if not all, their digital work, within collaborations.
As indicated by both the survey and interviews, these project teams are often inter-departmental and often span institutions, but not necessarily across national borders. The teams also tend to be either fairly small in terms of membership with either three to five members or larger with more than ten individuals involved. The team’s budgets were diverse with some (11%) having budgets of less than $50,000 (CDN) and others (37%) having larger budgets of more than $250,000 (Cdn). The rest of the projects had budgets between $50,000 and $250,000. The majority (66%) of respondents indicated that they had led a digital project in the past year. The rest of the respondents indicated that they were members within a team, provided technical support, or acted as project manager.

Survey and interview results suggest certain types of communication patterns for these teams. Survey respondents indicated that they relied on email and face to face meetings for most of their communication. There is relatively little use of online collaboration tools. Interestingly, given the digital nature of the work, the respondents generally rated face to face meetings as being “very effective” at a higher rate than they did their email correspondence. This finding is reinforced in the interviews where interviewees reported a high use of email for communication, a tool which is particularly useful for record keeping. However, they felt that face to face communication was more important for project planning and team development. One interviewee mentioned that one project in which they were involved “was going sideways because the team was not meeting enough.” This situation was corrected when the team began meeting on a regular schedule.

The teams varied in their use of formal documents that outlined roles, responsibilities, decision-making methods, and conflict resolution mechanisms. In terms of decision making, survey respondents evaluated the importance of various decision making mechanisms to their particular research team. Reflecting on their experiences, 79% respondents found decision-making by the project leader as important or very important to the team as compared to consensus (73%) or decision making by a few select individuals (66%). These results might reflect the large number of self-identified team leaders in the survey sample. In addition, most of the digital teams did not have a formal document outlining the organizational structure. Only 38% indicated that they did. Of this subset, 74% indicated that they felt that the organizational structure was well understood and effective.

The survey respondents were also asked about a team charter or formal document that outlines roles and responsibilities of team members. Again, like the formal document outlining organization structure, only 30% of respondents indicated that their team had one. Of this subset, 70% agreed or strongly agreed that the role and responsibilities were well understood.

Finally, research teams did not generally have formal documents outlining decision-making processes or dispute resolution. Only a small fraction indicated that their digital projects team had either in place. However, when these documents were negotiated, they were generally viewed as well understood and effective.

The opposite situation is true for formal project plan/project management documents that outline project goals, outcomes and timelines. In this case, 65% of respondents indicated that their team had such a document. Of this subset, 69% agreed or strongly agreed that processes and steps required to meet the project goals were well understood. In addition, 80% agreed or strongly agreed that the project goals were well understood. Finally, 64% agreed or strongly agreed that the mechanism for setting and sharing project goals was effective.

Despite the lack of formal documentation, the respondents reported that they were satisfied or very satisfied with the organization and processes of their digital projects team. In terms of project success, 71% felt that their digital project was successful or very successful. Notably, approximately, 12% felt that their project was neither successful nor unsuccessful. A small number (6%) reported that their digital project was very unsuccessful.

While survey respondents did not indicate directly reasons for success in their projects, they did outline challenges associated with team projects as well as their benefits through a series of open-ended questions. The interviews begin to provide some insight into these areas as well.

Both survey respondents and interviewees outlined challenges associated with digital projects. These include:

- Scheduling (e.g. coordination, organizing and scheduling face-to-face time)
- Time commitments (e.g. different workflows, planning the time needed for tasks correctly)
- Communication (e.g. geographical dispersion of team members, avoiding work duplication, keeping everyone on-board)
- Interpersonal (e.g. conflict resolution, incorporating different work rhythms and styles, not interfering)
- Expectation management (e.g. actually getting done what needs to get done, making expectations clear for each member, sharing responsibilities and goal attainment)
- Combining different skill sets and academic perspectives (e.g. lack of appreciation of the value of different contributions, need to think of work through another kind of specialist’s eyes, different vocabularies, different values of outcomes)
- Logistics (e.g. ensuring that resources are available)
- Need to build specific “team” skills (e.g. need to develop a working style and management skills that are often lacking in the environment, managing without proximity and formal lines of reporting, creating an environment where all voices, regardless of traditional authority, are respected and encouraged.)

As one of the interviewees stated succinctly, “team work is hard.” Among important skills highlighted were flexibility and patience. As one particular challenge, both survey respondents and interviewees also highlighted challenges associated with the culture and language gaps that might exist between team members from different backgrounds. One participant stated that “Humanities scholars do not know a lot about computer science or technology. Computer science individuals do not know about humanities. Information scientists know about data, but not programming.”

As a result, when representatives from these different groups work together, time must be spent ensuring that they understand each other and their respective roles within the digital project.
Building on the discussion of challenges, interviewees also explored factors that contributed to project success (or in some cases, non-success.) Generally, digital projects are perceived to be successful when the project outcomes are met while the team is still able to maintain strong interpersonal relationships. Project success appears more likely when every member of the team has a stake in the project. However, this stake does not necessarily have to be equal among team members. The interviewees also stressed that team members must have the skills and knowledge necessary to undertake the required tasks as well as understand their respective place within the project and the project’s goals and outcomes.

While not addressed specifically within the survey, the interviewees discussed the role and impact of training in team success. The interviewees indicated they generally had received very little formal training in team skills, such as communication, negotiation, conflict resolution and others. Instead, most learned to work in teams by doing in teams related to student project or those in other settings. This reality reinforced the importance to many about the need for good people with a team project, because not everyone may be able work as part of a team. One interviewee commented that they “were very choosy about whom they worked with”, especially after some less than pleasant experiences. Another survey respondent echoed this by suggesting that one should “make sure you’ve got good people; good people make a good team make good projects. Everything else is secondary.” In addition, several interviewees also commented that collaboration was a “state of mind.” In other words, individuals working within teams need to be open to new ideas, flexible, and ultimately patient.

By way of summary and conclusion to the survey and interviews, participants were asked for advice for teams undertaking their first digital project. The advice given echoes the discussion of benefits and challenges of team projects. In particular, the importance of organization was reinforced. As one survey respondent stated, “the point though is to be as organized as possible and to document as much as possible.” Planning plays an important role in this regard. Here, survey respondents suggested that one should “plan as much as you can before you begin” and to “set up a structure and deadlines with milestones.” The comments also focused on ensuring that “goals and expectations are clear from the beginning” and that each team member must understand their role within the project. In terms of developing working relationships, the respondents advocated “enough F2F (face to face) time and opportunities to build trust and working relationships before relying heavily on electronic methods of interaction and communication.”

5. DISCUSSION
This research contributes to a larger discussion regarding the nature of project teams within an academic setting. In an article examining academic-practitioner collaboration in management research, Amabile et al. [1] suggest it is necessary to understand the nature of collaboration and those factors which contribute to its success while minimizing the potential difficulties. Kraut et al. [12] and Suchman and Trigg [29] also suggest that team processes need to be understood in order to develop appropriate software tools to support them. Building upon this, Borgman [5] suggests that through understanding the nature of collaboration between scholars, better tools, services, and policies for information access and sharing can be created.

Digital projects are being undertaken by teams of individuals with a variety of skills, academic disciplines, and perspectives. In addition, people involved in this work are engaged in multiple projects at one time, meaning that they cannot commit 100% of their time to a single project, thus adding to the challenges. They are finding benefits and advantages to this approach to digital work, despite the associated challenges. Granted, as identified in the survey and interviews, digital projects demand a team approach given the variety of skills, content and methodologies that are required. Finally, the respondents and interviewees feel that their projects have been primarily successful.

Given the survey and interview results, several conclusions regarding work patterns of these teams are worth highlighting. First, these teams appear to be operating with relatively little formal documentation of roles, responsibilities, structure, and decision-making process. The exception is with documents that outline project goals, outcomes, and timelines. This may not be surprising given that nearly 50% of the teams in which the survey respondents are involved are comprised of three to five members and are predominantly based at one institution. The challenges that may flow from little formal document may become more apparent when project scope, membership and budget increase in size and complexity. For example, there are several very large digital projects being undertaken at present. Some of these have over 30 team members and budgets in the millions of dollars. The importance of documentation is also reinforced when considering the identified challenges and advice in the survey and interviews. The survey respondents and interviewees all focused on the importance of planning, scheduling, and organization.

Second, these results reinforce the value of different forms of communication within team work. While we feel comfortable in the digital world with email, wikis, and other forms of electronic communication, the importance of face-to-face communication cannot be underestimated. As the survey respondents and interviewees stressed, in certain contexts, face-to-face is more effective than email, particularly for planning, resolving difficulties, and addressing anything of ambiguous nature. Email, listservs, and others have their place, particularly as record keeping tools. These findings reinforce earlier studies undertaken in this field [24, 25].

Third, individuals involved in digital projects have, for the most part, learned team work and associated skills by doing it. This suggests a “trial and error” approach, which may not be the most efficient way of learning. Given the limited resources, including money, time, and staff, associated with many digital projects, this approach may become an inefficient use of resources. Having said this, opportunities may exist to capitalize on the learning that has already occurred by sharing it with other digital projects.

Finally, the survey respondents and interviewees, in particular, reinforce the importance of identifying the right people with whom to collaborate on digital projects. This suggests that while the technical components and content are important to a digital project, the project’s ultimate success (or not) may rest on interpersonal factors.
6. RECOMMENDATIONS AND CONCLUSIONS

The following recommendations are designed to support the already strong team work processes that are in place.

First, teams should ensure that they have strong communication processes in place, both face-to-face and electronic. In particular, teams might schedule regular face-to-face meetings, the frequency of which will be dictated by a specific project. As stressed in the interviews, these meetings are not only important for sharing information and making decisions, but also for reinforcing the sense of the team and the nature of the joint undertaking. In addition, teams should ensure that they build the cost of these types of meetings into funding applications. Teams that encompass a variety of different skill sets and academic perspectives might also benefit from a “translator” who can help navigate the language and culture gap between the various perspectives. For example, this translator might be an individual with both humanities and computer training and who can navigate between these two distinct communities.

Second, teams, particularly larger ones in terms of budget, membership size, and project scope, might consider more formal documentation of team and project structure and plans, decision-making processes, and dispute resolution mechanisms. This move supports the suggestion of one survey respondent to plan as much as possible in advance of undertaking the project work. The ultimate goal of the documentation is to maximize the benefits of collaboration while reducing the associated challenges. The process of determining the formal processes should be determined in the early stages of working together.

Third, this community might consider more formal training directed to the particular needs of digital project teams. This training would move beyond content and methodology and include courses and workshops in project management, communication, negotiation, problem solving and others [1, 6, 20]. There is a growing realization that collaboration requires new skills on the part of the researchers since a team works differently than an individual [2, 3, 9, 10, 12, 28].

An opportunity also exists to incorporate this skill development more formally into course curriculum. Besides training students in particular content and methodology, academic programs play a role preparing students for employment both within and outside the academy. Given that students are likely to work within teams after graduation, steps should be taken in advance to prepare them accordingly [20]. Given that students play an active role in the digital projects already underway, the project leaders could also use these opportunities to develop related team skills in the students, much in keeping with the “learning by doing” model already in place.

Finally, given the present “learning by doing” model, more team and self reflection could be incorporated to ensure active learning from the projects. At the end of any digital projects, teams could engage in a reflection process to explore the factors that contributed to project success and those that did not. It would also be an opportunity to evaluate the performance of team members and determine potential partners for future collaborations. As highlighted in the introduction, several digital teams have undertaken this activity and transferred knowledge to subsequent team endeavours [See, for example: 22]. Individual team members could also reflect on their own performance and determine if they are developing the appropriate collaborative “state of mind”, in the words of one of the survey respondents. As highlighted in the interviews and survey results, the “right” people for collaborative digital projects are those who are able to see the value in other perspectives and able to capitalize on the many benefits associated with digital team work.

Research teams are widely used to undertake various digital projects. The teams in which these participants have been involved have been successful and found ways to manage many of the various challenges associated with this type of work. This study is one step towards understanding the nature of these research teams while recommending several best practices.

7. REFERENCES


[31] Unsworth, J., Learning From NORA: Distributed Software Development in the Humanities. 2007: Bloomington, IN.

ABSTRACT
This paper builds on our prior research in which a Documentation Evaluation Model (DEM) for social science data was constructed and hypotheses about impacting factors of perceived documentation quality were proposed. In this paper, results from interviews and a survey were used to validate the model and test those hypotheses. We found the DEM model is valid, and that perceived documentation quality varies with several characteristics of data and is weakly affected by users’ absorptive capacity.

Categories and Subject Descriptors
H.3.7 [Information Storage and Retrieval]: Digital Libraries

Keywords
perceived documentation quality, secondary data use, social science data, absorptive capacity.

1. INTRODUCTION
In this paper we present the results of an empirical test of a Documentation Evaluation Model for social science data. In previous work, we presented the Documentation Evaluation Model (DEM), which evaluates perceived documentation quality on two dimensions: sufficiency and ease-of-use [3]. There we introduced the context of data sharing in social science research, the concept of secondary data analysis, and the importance of documentation as a knowledge transfer channel between data producers and secondary users. Secondary data analysis is the analysis of data collected by other people. Documentation refers to knowledge about data that is recorded and transferred to secondary users and that helps users understand and use the data. We also proposed several hypotheses about what affects documentation sufficiency. The proposed impacting factors of perceived documentation sufficiency can be grouped into two categories: users’ absorptive capacity and characteristics of data. Absorptive capacity means users’ general ability to understand and use data. This paper presents findings from a survey and interviews of secondary users of social science data that we used to validate DEM and test hypotheses about the factors that affect documentation quality.

2. METHODOLOGY
We designed a survey targeted at researchers who conduct secondary data analysis in social science research. The objective of the survey was to gather data from secondary users about their experiences using data and documentation and their perceptions about the sufficiency and ease-of-use of documentation. We built a distribution list for the survey from the following sources: (1) respondents to a related study who were willing to be contacted for a further study; (2) the mailing list of the International Association of Social Science Information Services and Technology (IASSIST); (3) a list of authors of publications based on secondary data analysis compiled by searching research literature databases; (4) publication data from data archive websites, such as the Interuniversity Consortium for Political and Social Research (ICPSR) and the Economic and Social Data Service (ESDS) of the UK, and names of individuals who are liaisons between these data archives and data users, and (5) names of additional secondary data users from other sources (snowball method).

Part of the survey instrument was pre-tested in a related study. We pre-tested the entire survey instrument again by observing and interviewing two graduate students at the University of Michigan. The formal survey started in May 2008. We sent a total of 1,260 surveys, either directly to secondary data users, or by way of data librarians who we asked to forward our recruiting email to their users. By August 31, 2008, 431 surveys were received in total. After removing responses for instances where data had not been used for research, were used for research outside of social science, or where use occurred more than 10 years ago, there are 387 usable responses. The survey collected both highly structured quantitative data and open-ended qualitative data. We also interviewed 13 secondary data users for 30 to 70 minutes each. Secondary users were asked questions about their secondary data analysis experiences in general and their most recent experience with a particular dataset in unstructured exploratory interviews. Preliminary findings from the interviews informed the survey design and supplement the quantitative survey results. The unit of analysis for the survey is a single instance of using a single social science dataset and documentation. Each respondent was asked about his or her most recent experience using a dataset created by a different individual or entity for which the analysis was completed.
3. RESULTS

3.1 Reliability and Validity of DEM

The DEM model consists of two components for measuring perceived documentation quality: ease-of-use and sufficiency. There are strong correlations between sufficiency, ease-of-use, and an overall measure of perceived documentation quality. Cronbach’s alpha for sufficiency and ease-of-use as two sub-measures of documentation quality is 0.8669, which is much higher than the acceptable threshold 0.7. The strong correlation between the two measures also tells us more sufficient documentation tends to be easier to use.

Sufficiency and ease-of-use are average values of their sub-measures. Ease-of-use is measured by how much users agree with the following four statements: 1) overall the documentation is easy to use, 2) it is easy to learn to use the documentation, 3) it is easy to find information in the documentation, and 4) the content of the documentation is clear and understandable, all measured on a 7-point Likert scale. There are strong correlations between these sub-measures. The Cronbach alpha of ease-of-use is 0.9474. Factor analysis showed these four measures are loaded onto one factor. Sufficiency has four sub-measures. Completeness is measured by respondents’ ratings of how well each element is described in documentation, from a set of the most important elements (sampling frame, response rate, variable definition, etc.) that we selected from the Data Documentation Initiative (DDI), an international standard for documenting social science data. The other sub-measures are users’ rating of these statements: 1) the documentation provided enough information for me to judge the reliability of the data, 2) with the documentation, I did not need additional information about the data for my use, and 3) the documentation provided enough information for my purpose of use. The Cronbach alpha of sufficiency is 0.8011. Factor analysis showed these four measures are loaded onto one factor.

Qualitative data from the interviews shed more light on the validity of the model. Consistent with the DEM, we found that most of the problems users encountered with documentation fall into two categories: sufficiency and ease-of-use. Many users commented about the absence of certain elements or incomplete descriptions, which supports completeness as a valid measure. Some data have very limited documentation, such as when only variable names or a questionnaire are provided. Some documentation is very brief and tersely written. Examples of information missing from the documentation include the meaning, measurement, construction, and source of variables, methods used for recoded and derived variables, the process of collecting and entering data, sample design, response categories, response rates and anomalies in responses, instrument reliability, limitations of data, information about the purpose of the study, links between variables and the data collection instrument, links between different levels of data, the time period covered, details of experiment design, the reasons for missing data, imputation of missing data, the rationale for changing measures between waves, relationships between a dataset and other data collected by the same producer, coding categories, and skip patterns.

Respondents used the following terms to describe documentation that is hard-to-use: irritating, annoying, confusing, cumbersome, complicated, fragmented, dispersed, not user-oriented, not transparent, puzzling, long, huge, massive, unorganized. Also consistent with the DEM, problems related to ease-of-use can be categorized into the following categories: (1) Hard to find the information needed because documentation was available only in hard copy, information was dispersed across multiple files, no cross-references were provided between various parts of the documentation, an unorganized and overwhelming amount of information, or the documentation and data are stored in different places. (2) Hard to understand because descriptions were too tersely written, the terminology was not always clear, and scanned codebooks were blurry and very difficult to read. No respondents mentioned difficulties in learning to use documentation. Even though statistically the measure “easy to learn to use documentation” has high correlations with other measures, which is a sign of construct validity, we decide it was not necessary and should be dropped from the model. After removing “easy to learn”, the Cronbach’s alpha of ease-of-use is still 0.9380. The remaining three sub-measures are still loaded onto one factor.

Besides sufficiency and ease-of-use, users mentioned the accuracy of the documentation as another type of problem. Users often detect errors in documentation based on the inconsistencies between data and documentation. For example, questions of accuracy arise when response values do not match the skip patterns in the questionnaire, when weights in the data do not match weights in the documentation, and when the electronic and hard copy versions of the documentation are different. We anticipated this problem when we proposed the DEM. We decided not to include accuracy in the model because inaccuracy is often closely related with data, and we wanted to focus only on documentation quality. Accuracy needs to be added as an indicator for quality if a more complete evaluation model for both data and documentation is constructed.

3.2 Perceived Documentation Quality Varies With Characteristics of Data

Our interviews showed that perceptions of documentation quality differ between data produced for self-use versus data produced for sharing. Data for self-use are typically produced by individual researchers and small research groups or by government agencies for administrative purposes. Producers of data for self-use often do not have incentives to document data for secondary users. Data for sharing are typically produced by survey research organizations, commercial data companies, and some government agencies with a tradition of producing and sharing data, such as the Census Bureau, Bureau of Labor Statistics, and the Center for Disease Control. The mission of those producers is to collect high quality data for other researchers. Their performance is evaluated by how many users they attract and how much valuable research is conducted using their data. Those data producers are highly motivated to share their data, document data well, and provide user assistance. Statistics based on our survey support findings from the interviews. Data produced for self-use has documentation that is less sufficient (t = -3.6497, p = 0.0005), harder to use (t = -2.4644, p = 0.0077), and of poorer quality (t = -3.6562, p = 0.0002) than data produced for sharing.

Documentation for quantitative survey data and census data is more sufficient (t = 3.2798, p = 0.0007), easier to use (t = 2.2377, p = 0.0138) and of better overall quality (t = 2.7215, p = 0.0039) than that for administrative records, qualitative interviews and experiments. Documentation for qualitative data is perceived to be less sufficient (t = -1.6567, p = 0.0535), harder to use (t = -
with data producers than quantitative data users (chi² = 5.5555, p = 0.006). Qualitative data users are more likely to collaborate than users who obtained data from intermediaries (chi² = 9.3465, p = 0.018). Therefore it may not be that collaboration with data producers causes users to perceive that the documentation is of lower quality, but that users of data with low quality documentation tend to collaborate with data producers in order to use the data. Users who obtained data directly from producers are also more likely to collaborate with data producers than users who obtained data from intermediaries (chi² = 9.5555, p = 0.018). Therefore it may not be that collaboration with data producers causes users to perceive that the documentation is of lower quality, but that users of data with low quality documentation tend to collaborate with data producers in order to use the data. Users who obtained data directly from producers are also more likely to collaborate with data producers than users who obtained data from intermediaries (chi² = 9.3465, p = 0.002) such as data archives, data centers, and other secondary data users.

### 3.3 The Effect Of Users’ Absorptive Capacity On Perceived Documentation Quality

We borrowed the term “absorptive capacity” from Szulanski to refer to users’ ability to understand and use data [4]. Through interviews, we found users’ absorptive capacity for understanding and using secondary data is composed of three types of knowledge: knowledge about the data, background knowledge used to understand and interpret data, and data analysis skills. Below are some examples of the three types of knowledge.

- **Knowledge about data:** (e.g., what is the response rate and sampling frame for a particular survey, or how are the missing responses treated).
- **Background knowledge:** (e.g., disciplinary consensus on how to use common types of data, how to determine whether or not to weight variables from samples, which variable best captures certain concepts, how to interpret frequently occurring variables, and how to handle specific measurement issues).
- **Data analysis skills:** (e.g., how to convert hierarchical data files into appropriate rectangular files or how to construct new derived variables).

Borgman (p. 230) talked about the effects of knowledge distance between data producers and users. “The greater the distance from the data author, the more scientists must rely on documentary evidence. Scholars in an area farther away from the field in which the experiment was conducted, students new to the field, teachers, and policy makers will be constrained in their ability to interpret the available documentation of the context [1].” Zimmerman (2003) explained how users’ formal (academic training) and informal knowledge (field experiences) affect users’ ability to understand secondary data [5].

Based on the literature, we created two sub-measures for absorptive capacity: professional status and users’ rating of the statement “the topic(s) of the data were within my specialty.” Data from the interviews are mostly consistent with the literature, but also revealed more details of absorptive capacity.

1. Prior experience using the same data makes data using easier. “For me, it is easier to use datasets that are familiar. If I use serial data that comes out every year, that is easier for me, because I see it on a regular basis.”

2. Prior experience using the same data makes documentation easier to use. “It (the documentation) is kind of hard to look through. (But) since I am already very familiar with the data, it is easier for me to look through it.”

3. Prior experience using the same data makes users less reliant on documentation. “The most recent time I used the data, I didn’t use documentation, because I have used it so many times and I know many things in my head.”

4. Secondary data analysis experience builds data analysis skills. “I work with secondary data over and over again. I learn through the whole process how to be more efficiently collect the data in the first place, data manipulation, the creation of variables, and selection of variables. Through the process, I become more effective in interpreting results, identifying positive or significant results.”

5. More background knowledge has a negative impact on perceived trustworthiness of data. “the less you know, the more you are willing to step on faith. Less experienced Ph.D. students might trust the data more. . . . With more experience, you can use the data more easily. With less experience, you trust the data more.”

6. Data collection experience has a negative impact on perceived data and documentation quality. “I never doubted the quality of data before I did the fieldwork of collecting data for China Statistics Bureau. . . . I interviewed an illiterate poor farmer at northwest China. The questionnaire was long. He cared about the
incentive money and would like to answer my questions. At the beginning, he thought carefully before answering my questions, but then later, he became tired and impatient, and just answered without thinking.  

Based on the interview data, we added another four sub-measures of absorptive capacity: 1) users’ experience using exactly the same data, 2) experience using secondary data in general, 3) experience analyzing self-collected data, and 4) experience collecting data. It turns out the latter two measures are highly correlated (0.8409). Factor analysis found them loaded onto one factor. Cronbach’s alpha for the two sub-measures is 0.9121. We combined them into one sub-measure: experience with primary data. We then tested whether the remaining 5 sub-measures could be combined into one construct: absorptive capacity. Cronbach’s alpha for the remaining five measures is 0.5621, which is undesirable but acceptable. Factor analysis showed the five measures are loaded onto one factor, which explains 38% of the variability of data, greater than the threshold 25%. Factor loadings are moderate, higher than 0.4 but lower 0.8. These numbers showed that the five measures can be combined into one factor, but the relationships are not very strong. To ensure the validity of our results, we analyzed the both the relationships between the combined factor absorptive capacity and dependent variables, and the relationships of each sub-measure and dependent variables.

For data produced for self-use, there is a non-significant negative correlation between users’ data collection experience and sufficiency for reliability judgment (r = -0.03, p = 0.8), and a non-significant negative correlation between users’ experience with primary data and reliability judgment (r = -0.04, p = 0.73). Those non-significant negative correlations become significant and positive for data produced for sharing (r = 0.1858, p = 0.0039, r = 0.2020, p = 0.0016). We know from the previous section that data produced for sharing tend to be better documented than data produced for self-use. These numbers can be interpreted this way: for poor documentation, experience in collecting and analyzing primary data does not affect users perception about how sufficient the documentation is for judging the reliability of data. Put it in a simple way, poor documentation is regarded as poor no matter how experienced the users are. Good documentation is more sufficient for experienced users than novice users. These statistics are slightly different from the aforementioned findings from interviews, but are more convincing because they are based on a much larger sample.

The relationship between absorptive capacity and perceived documentation quality is small but significant. Correlation between absorptive capacity and sufficiency is 0.2652 (p = 0), between absorptive capacity and ease-of-use is 0.1798 (p = 0.0006), between absorptive capacity and overall quality is 0.1359 (p = 0.0099). The correlation of each sub-measure of absorptive capacity and each aspect of documentation quality is listed in the following table. Each of the five measures has small but significant relationships with perceived sufficiency. Experience analyzing the same data, secondary data analysis experience in general and familiarity with the topics of data have small but significant effects on ease-of-use and the overall measure of good documentation.

### Table 1. Correlations between sub-measures of Absorptive Capacity and Perceived Documentation Quality.

<table>
<thead>
<tr>
<th></th>
<th>Sufficiency</th>
<th>Ease-of-use</th>
<th>Well documented</th>
</tr>
</thead>
<tbody>
<tr>
<td>Professional status</td>
<td>0.15*</td>
<td>0.07</td>
<td>0.07</td>
</tr>
<tr>
<td>Experience with the same data</td>
<td>0.29*</td>
<td>0.22*</td>
<td>0.17*</td>
</tr>
<tr>
<td>Secondary data use experience</td>
<td>0.2*</td>
<td>0.17*</td>
<td>0.14*</td>
</tr>
<tr>
<td>Data topics with users' specialty</td>
<td>0.22*</td>
<td>0.19*</td>
<td>0.19*</td>
</tr>
<tr>
<td>Experience with primary data</td>
<td>0.12*</td>
<td>0.04</td>
<td>0.0015</td>
</tr>
</tbody>
</table>

3.4 The Effect Of Documentation Quality On Incentives To Use Secondary Data.

Inadequate documentation entails high use costs and possibly could lead to incorrect conclusions. Secondary users of datasets produced by small producers often say they prefer not to use other people’s data. For example, a health psychologist complained about poor documentation of a dataset collected by another researcher. The only documentation provided was variable names. She did not know what types of questions were asked, how the questions were asked, or how the questionnaires were administered. It was hard to differentiate similar variables. She described the use process as a hassle. She would have preferred to collect her own data, but she used the data because she was assigned to do so. Interviewees described administrative records as messy and poorly documented. For example, a user mentioned his difficulty using a set of police records where police officers used the address of the parking lot at the police station for the residential address of homeless prisoners without including this coding practice in the documentation. When a secondary user analyzes the data and tries to identify hot spots for crime, the police station appears to be a hotspot. If the user trusts the data without further investigation, wrong conclusions can be drawn. Interviewees still would use administrative records because there is no other way to get the data.

Documentation of data produced for sharing is good in general but not perfect. Even when data is produced for sharing, many users still seek additional information beyond what is provided in the documentation. Some users reported that the documentation of data produced for sharing is massive and hard to use, but no users

---

2 As measured by respondents’ rating of the statement: “the documentation provided enough information for me to judge the reliability of the data.”

3 Here the experience means users’ experience in collecting and analyzing primary data.

* p<0.05.
said they would not use the data because documentation is not ideal. In fact, users expressed appreciation, understanding, and forgiveness for problems with documentation for datasets that are produced for sharing. One user said: “This dataset is extremely extensive and complex by nature. The documentation was very good given the data complexity.”

Many researchers rely almost exclusively on large-scale data for their research. For them, using secondary data produced for sharing is necessary and routine because they cannot collect data themselves on such a large scale. One user said: “I primarily use secondary data. Those secondary data allow me to look at information at macroscopic scale that I wouldn’t be able to do otherwise. If I collect the data myself, I can ask the specific question to which I am looking to answer. But the data would not be as representative and rigorous as the data collected by professional producers.” Some users of this kind of data are willing to take on the additional burdens associated with using large complex datasets. They think it is natural that documentation does not provide everything. It is part of their research process to interact with other researchers for secondary data use. Some users understand that producing very high quality documentation is time consuming, and may mean that release of data would be delayed if the documentation had to be ideal. Facing a choice between waiting longer and perfect documentation, they prefer data to be released sooner.

4. CONCLUSIONS
This paper is a part of a larger study, which studies perceived documentation quality of social science data and the information seeking of secondary data users to overcome inadequate documentation. In this paper, we tested a Documentation Evaluation Model and the impacting factors of perceived documentation quality proposed in a previous publication. We found that DEM is a valid model for perceived documentation quality, and that perceived documentation quality varies with the characteristics of data and is weakly affected by users’ absorptive capacity. We also reported some findings about how perceived documentation quality affects users’ incentives to use secondary data. While poor documentation entails high use costs and affects users’ incentive to use data produced by small producers, problems with documentation do not have a significant impact on users’ incentives to use administrative records and data produced for sharing. In the future, we will report findings about how users overcome inadequate documentation by seeking information beyond what was provided by documentation and whether their information seeking behavior varies with characteristics of data and users’ absorptive capacity.

5. ACKNOWLEDGMENTS
This research is funded by the National Science Foundation, Award # IIS 0456022 as part of the project entitled “Incentives for Data Producers to Create Archive-Ready Data Sets.”

6. REFERENCES
Data Access and Long-Term Data & Knowledge Preservation for Earth Science: An Overview on Some ESA Initiatives

Sergio Albani†, Vincenzo Beruti†, Luigi Fusco†, David Giaretta‡

†European Space Agency
ESRIN, Via Galileo Galilei
00044 Frascati, Italy
{Vincenzo.Beruti, Sergio.Albani, Luigi.Fusco}@esa.int

‡STFC
Rutherford Appleton Lab
Didcot, Oxon, OX11 0QX UK
david.giaretta@stfc.ac.uk

ABSTRACT

Earth Observation Space Missions provide continuous surveillance of the Earth producing huge amounts of data every year that need to be processed, elaborated, appraised and archived by dedicated systems.

The ability to preserve these data in the long term and to provide easy access in order to facilitate their exploitation and utilisation is a fundamental issue and a major challenge at programmatic, technological and operational levels.

This contribution will provide an overview on some ESA initiatives carried out in collaboration with European entities and institutions, with the objective of guaranteeing long term data preservation.

In particular the paper will focus on the ESA participation and contribution in the CASPAR project, the PARSE.Insight project and the Alliance for Permanent Access to the Records of Science coalition.

Categories and Subject Descriptors

H1.1 [Models and principles]: Systems and Information Theory
H3 [Information storage and retrieval]
K.6.1 [Management of Computing and Information Systems]: Strategic information systems planning

General Terms

Management, Documentation, Standardization, Languages, Theory, Legal Aspects.

Keywords

Digital preservation, digital curation, infrastructure, workflows.

1. INTRODUCTION

For the implementation and the systematic monitoring of international environmental conventions, the Earth Science communities need access to vast amounts of Earth Observation (EO) and Earth Science data to derive objective information and to share knowledge in all environmental sensitive domains over a continuum of timescales (from historical measurement to real time assessment to short and long term predictions) and a variety of geographical scales (from global scale to very small scale).

Access to full historical data collections, including the raw and processed data, scientific analyses, models and results, needs therefore to be guaranteed.

Locating and accessing historical data is a difficult process and their interpretation can be even more complicated given the fact that scientists may not have (or may not have access to) the right knowledge to interpret these data. Storing such information together with the data and ensuring all remain accessible over time would allow not only for a better interpretation but would also support the process of data discovery, now and in the future.

An overall strategy for the preservation and access to such data is however not in place despite the fact that most mission operators identify long-term EO data archiving as an important issue where cooperation and sharing can bring a benefit. So there is an urgent need for a coordinated and coherent approach for a harmonized European archives management policy; the European Space Agency (ESA) supports and is involved in various initiatives to improve EO space data preservation, access and exploitation.

ESA has, in the very recent past, initiated an important operational initiative to ensure and facilitate the accessibility and usability of the preserved European EO space data set through the implementation of a cooperating / harmonized distributed solution (European Long Term Data Preservation System) based on the application of a European LTDP Common Policy and sustained through cooperative (multi-source) long term funding schemes.

The integration of emerging technologies such as the GRID and digital libraries, the involvement in dedicated coordination activities and RTD projects funded by the EC (e.g. CASPAR and PARSE.Insight) and the development of operational solutions (as the OAIS compliant Standard Archive Format for Europe, SAFE, format and Multi-Mission Facility Infrastructure, MMFI, for EO data management) are considered by ESA as fundamental
activities for improving data access and long-term data and knowledge preservation in the wide Earth Science distributed community.

This contribution provides an overview on some important ESA initiatives carried out in collaboration with other European entities and institutions, with the objective to guarantee long term Earth Science data preservation. In particular the paper focuses on the ESA participation and contribution to the CASPAR and PARSE.Insight projects and to the Alliance for Permanent Access to the Records of Science coalition.

2. ESA INITIATIVES
2.1 The CASPAR Project

CASPAR - Cultural, Artistic and Scientific knowledge for Preservation, Access and Retrieval - is an Integrated Project co-financed by the European Union within the Sixth Framework Programme (Priority objective IST-2005-2.5.10, "Access to and preservation of cultural and scientific resources").

CASPAR [1,2] addresses the growing challenge facing society of a deluge of intrinsically fragile digital information, upon which it is increasingly dependent, by building a pioneering framework to support the end-to-end preservation "lifecycle" for scientific, artistic and cultural information, based on existing and emerging standards.

The ambitious challenge to build up a common preservation framework for heterogeneous data and a variety of innovative applications will be achieved through the following high level objectives:

- to establish the foundation methodology applicable to a very wide range of preservation issues (the guiding principle of CASPAR is the application of the OAIS Reference Model);
- to research, develop and integrate advanced components to be used in a wide range of preservation activities (these components will be the building blocks of the CASPAR Framework);
- to create the CASPAR framework, the software platform that enables the building of services and applications that can be adapted to multiple areas.

The CASPAR consortium of 17 European partners (8 institutional such as ESA and UNESCO, 3 universities and 6 private companies) will demonstrate the validity of the CASPAR framework through heterogeneous testbeds, covering a wide range of disciplines from science to culture to contemporary arts and multi-media, providing a reliable common infrastructure which can be used or replicated in many more areas.

The CASPAR consortium is seeking to guarantee the future evolution of CASPAR. This ambitious goal is being pursued through:

- the building of the CASPAR preservation user community creating consensus around the initiative and gathering a critical mass of potential users;
- embedding the CASPAR framework and components within key memory organisations, both national and international.

In CASPAR, ESA plays the role of both user and infrastructure provider for the scientific data testbed using new proposed standards (e.g. SAFE, developed by ESA in the framework of its Earth Observation ground segment activities), specialized infrastructures (e.g. MMFI, a distributed system based on the OAIS standard that provides producer oriented services for data archiving, data retrieval and processing management) and open GRID environment.

ESA participation to the CASPAR project is driven by:

- the interest in confirming and consolidating the validity of the OAIS reference model, already adopted in several internal initiatives, for long term preservation issues.
- the interest in developing preservation techniques covering not only the data but also the knowledge associated to them. In fact it is crucial to guarantee the ability for different user communities to understand the meaning of the data in order to use them. This will be possible by generating and preserving appropriate Representation Information that will be returned with the data to different users in the quantity and form most suitable for their knowledge base and needs.

The ESA selected scientific dataset consists of data from GOME (Global Ozone Monitoring Experiment), a sensor on board ESA ERS-2 (European Remote Sensing) satellite, in operation for more than a decade.

In particular, the GOME dataset has a large total amount of information distributed with a high level of complexity, is unique because it provides more than 11 years global coverage and is very important for the scientific community and the Principal Investigators (PI) that on a routine basis receive GOME data (e.g. KNMI and DLR) for their research projects (e.g. concerning ozone depletion or climate change). Note that GOME is just a demonstration case because similar issues involve many other Earth Observation instrument datasets.

The GOME data set includes different data products, processing levels and associated information. The commonly used names and descriptions of these types of data are as follows:

- Level 0 - raw data as acquired from the satellite, which is processed to:
  - Level 1 – which gives measures of radiances/reflectances. Further processing of this gives:
    - Level 2 – which provides geophysical data as trace gas amounts. These can be combined as:
      - Level 3 – which consists of a mosaic composed by several level 2 data products with interpolation of data values to fill the satellite gaps.

In addition these are a number of types associated pieces of information which must be available.

- Level processors (needed to process data from one level to another) and format converters,
- Auxiliary data (the ancillary info needed to process data),
- Documents and methods,
- Data viewers (e.g. SAFE Toolbox),
Examples of GOME science applications.

As example of complexity of digital objects to be integrated in the CASPAR testbed, Figure 1 illustrates the processing scheme to derive GOME Level 2 data from Level 0.

The core of the CASPAR dedicated testbed is the preservation of the ability to process data from one level to another, that is the preservation of GOME Level 0 data and of all components (processors, metadata and ancillary data, knowledge and documentation, ontology, algorithms and their evolution and validation history) that enables the operational processing for generating products at higher levels.

As first demonstration case, it has been decided to preserve the ability to produce a GOME Level 1C data (fully calibrated data) starting from a Level 1B data (raw signals plus calibration data). Because a single Level 1B data can generate several Level 1C products (applying different calibrations), the archive will store Level 1B data, the processor needed to generate Level 1C and all the information needed to perform this process depending on the user needs and knowledge.

By the end of the CASPAR project (planned for the second half of 2009) the ESA testbed intends to demonstrate the preservation of this GOME processing against changes in hardware, software, environment and Designated Community knowledge-base. It must cope with changes related to the processing by managing correct information flow through the system, the system administrators and the users, using the CASPAR components. These include an alerting system, the ability to return appropriate Representation Information to users with different knowledge bases, the ability of the system administrators to access the source code of the processor in order to upgrade it and all the representation information related to it, a notification system in order to ensure that the different users communities will be correctly notified of any changes.

Figure 1. GOME Level 0 to Level 2 processing
2.2 PARSE.Insight project

PARSE.Insight - Insight into issues of Permanent Access to the Records of Science in Europe - is a project in the topic area "INFRA-2007-3.3: Studies, conferences and coordination actions supporting policy development, including international cooperation, for e-Infrastructures" [3].

The FP7 e-Infrastructures initiative has already resulted in a number of projects being set-up or in negotiation covering a range of areas of scientific endeavor. The creation of virtual research communities brings together stakeholders with diverse backgrounds and capabilities. One consequence is a large diversity of approaches and there is as yet little consensus across these communities about best practices in maintaining permanent access to the records of science in a sustainable way.

The purpose of this project, started in early 2008, is to work with providers, publishers and users of scientific information and large repositories to deliver insight into the issues of sustainable permanent access and to provide requirements and cross-fertilisation of ideas 1) between providers and users, 2) between the various sectors of interest they represent, and 3) between the research community as a whole and national/European funding agencies.

PARSE.Insight considers the following six key outcomes:

- Generate a roadmap for the support e-infrastructure for long-term accessibility and usability of scientific and other digital information in Europe;
- Provide insight into current and planned research and development relating to e-infrastructures and permanent access, national, European and global, regardless of the funding mechanism;
- Identify gaps in the existing and planned infrastructure and the research agenda;
- Analyse the ability to share and capitalise on best practices as well as understanding the impact that e-Science is having on the research communities that it is serving, through an insight study into the capabilities and practices within the various research communities;
- Ensure better-informed investment decisions and sustainable e-repositories through an impact analysis framework and tool based on these insights;
- Initiate an international process for evaluating the sustainability and trustworthiness of digital repositories, and identifying best practice.

During the 24 months duration of PARSE.Insight, ESA will play a major role in all the activities above and in particular in the one that aims (through a research and services community overview, targeted surveys and in-depth case studies) to provide information that is needed to perform the gap analysis and refine the roadmap. Moreover ESA will lead the workpackage that aims not only to identify best practice among existing digital repositories but also to reach a common understanding of mechanisms to evaluate the sustainability and trustworthiness of e-infrastructure repositories.

2.3 The European Alliance for Permanent Access to the Record of Science

The Alliance is a unique cross-sectoral coalition of major stakeholders in science and scientific information, created in response to the work of a high-level task force initiated during the Dutch Presidency of the EU in 2004 [4].

The Alliance aims to provide a conceptual and strategic framework that organises and consolidates the individual efforts that the members of the consortium spend on various long-term preservation initiatives and fills possible gaps between them. This should ensure a unified, Europe-wide approach to the issue whilst working closely with relevant projects undertaken outside Europe.

ESA is one of the participants to the consortium, that currently include among others CERN, STFC, MaxPlanckG, CNRS, ESF, JISC, British Library, Koninklijke Bibliotheek, Deutsche National Bibliothek, Association of Scientific Technologic and Medical Publishers, National Preservation coalitions (UK, Germany, Netherlands).

At a practical level, Alliance members are already active in a number of relevant key projects for long-term preservation and access such as CASPAR, DRIVER, PLANETS, SHAMAN and DPE. The involvement of commercial partners in some of those projects, particular those in the storage and ICT industries, such as IBM, Microsoft and SUN, further expands the sphere of influence of the Alliance and its members.

Members are also participating in development of a new generation of scientific digital repositories which takes fully in account the European effort for building e-infrastructures such as GENESI-DR and EuroVO-AIDA, covering Earth Science and Astrophysics communities (both projects funded in the first FP7 Research Infrastructure call).

The Alliance therefore has adopted a strategic work plan, closely related to the CASPAR approach, to help consolidate its role at national, European and international levels. The work plan is a framework for actions that will help generate critical mass and accelerate the progress that the Alliance can make over the mid-term (2008-2010), which will help to multiply the investments made by the EU many-fold by helping to align the investments within individual nations.

3. ESA LONG TERM EARTH OBSERVATION DATA PRESERVATION AND CONCLUSIONS

The current ESA strategy for long term EO data preservation is based on the assumption that there is a fundamental requirement to guarantee to the scientific and operational user communities access and use for as long as possible to long time series of EO data for long term scientific research and environmental monitoring.

Work in the CASPAR project is providing prototypes for the preservation components and techniques needed. These are supplemented by the policy views of PARSE.Insight and should be taken up, in the medium to long-term, in the shared infrastructure which the Alliance for Permanent Access will bring together.
This contribution has provided an overview on some ESA-ESRIN initiatives carried out in collaboration with European entities and institutions, with the objective to guarantee long term data preservation. In particular the paper focused on the ESA participation and contribution to the CASPAR project, the PARSE project and the European Alliance for Permanent Access to the Record of Science coalition.

ESA participation to such initiatives will represent a big step for a coordinated and coherent approach for a harmonized European archives management policy.

4. REFERENCES


Speech Acts and Electronic Records

William E. Underwood
Georgia Tech Research Institute
Atlanta, GA 30332 USA
william.underwood@gtri.gatech.edu

ABSTRACT
All written records of human activity involve actions expressed in records. Archivists identify these acts when they describe records and when they review records for possible restrictions on disclosure. This paper reports the results of an analysis of Presidential records to determine the speech acts conveyed by the records and the role of these speech acts in describing the records. It also proposes a method for automatically recognizing these acts for use in support of archival description and review.

Categories and Subject Descriptors
H.3.7 [Information Storage and Retrieval]: Digital Libraries

Keywords
speech acts, performative verbs, pronominal co-reference, e-records, archival description, archival review

1. INTRODUCTION
Intellectual control of an archival collection is not achieved until the collection has been fully described. Furthermore, a collection cannot be fully described until archivists have read, understood, and recorded details as to the document types, the creators, the titles, dates, extent and contents of the records in the collection. Archivists and researchers will not have the capability to reliably locate relevant records and understand those records until catalogs, finding aids, and indexes are produced.

Given the increasing volumes of Federal and Presidential electronic records acquired by NARA, it will be decades, if not centuries, before full intellectual control of these records is achieved. This is especially true of Presidential e-records as they must be reviewed page-by-page for restrictions on disclosure as provided by the Presidential Records Act (PRA). Techniques for automatic description of records, file units and record series could provide archivists with earlier intellectual control of record collections.

The Freedom of Information Act provides that citizens may request Presidential records 5-years after the end of an administration. Archivists at Presidential Libraries do what they can during that 5-year period to gain a modicum of intellectual control that will enable them to effectively respond to FOIA requests. To respond to FOIA requests that include e-records, full-text search of collections, even with Boolean queries and automatic description of records, file units and record series could provide archivists with earlier intellectual control of record collections.

The purpose of this paper is to summarize progress on the investigation of the problem of recognizing speech acts conveyed by Presidential e-records and the roles of these speech acts in record (or item) description [1]. The investigations of the problems of document type recognition and topic recognition as they relate to record description are described in other reports [2, 3].

In section 2, the role that actions play in records and in archival description and review is discussed and the concepts of speech act theory are summarized. Section 3 summarizes the results of an analysis of Presidential records to identify the speech acts conveyed by those records. Section 4 describes a test of the thesis that speech acts identified in a record support archival description of the record. In section 5, a method is proposed for automatically identifying the speech acts in e-records. In section 6, the results of this paper are summarized.

2. RECORDS AND SPEECH ACTS
2.1 Records and Actions
It is a basic principle of Archival Science that to be a record of an activity, a record must be associated with some action [4]. The record must:

This work is licensed under the Creative Commons Attribution-Noncommercial-No Derivative Works 3.0 Unported license. You are free to share this work (copy, distribute and transmit) under the following conditions: attribution, non-commercial, and no derivative works. To view a copy of this license, visit http://creativecommons.org/licenses/by-nc-nd/3.0/.

DigCCurr2009, April 1-3, 2009, Chapel Hill, NC, USA
1. Carry out an action, e.g., a request that someone do something, a recommendation that someone do something, an appointment to an office (a dispositive record),

2. Provide evidence of an action already carried out, e.g., a report of actions taken (a probative record),

3. Provide information on which to base action, e.g., a biography of a candidate for federal office (a supporting record), or

4. Provide information that is about acts but is irrelevant to performing an action, e.g., a statement by the Press Secretary (a narrative record).

Actually, each of these types of record carries out an act. The dispositive records, by definition, carry out an act, e.g., request, recommend, appoint. A probative record also performs an action, e.g., a report of actions taken carries out the action of reporting an action. A supporting record also performs an act, e.g., the biography of a candidate for federal office is a statement of facts. A narrative record, e.g., a statement by the Press Secretary, is a statement of facts and thus carries on an action.

2.2 Actions and Record Descriptions
Archivists at Presidential Libraries title collections, record series, file units (folders), and individual items (records). They have been able to accomplish the titling of collections, record series and file units within the first five years of acquiring the records of a Presidential administration. After review of the records in a record series, archivists summarize the contents of the series in scope and content notes. If records are included in NARA’s online Archival Research Catalog, the individual items must also be titled. Due to the increasing volume of Presidential records, and due to the fact that archivists at Presidential Libraries must respond to Freedom of Information Act requests beginning five years after the end of an administration, it will be centuries before the holdings of a Presidential Library are reviewed and described.

NARA’s Lifecycle Data Requirements Guide is one of the standard guides for archival description. It prescribes that the name of a record’s document type should be included in the title. If the record is created by a particular person or position in an organization, then that person’s name or position title should also be included in the title of the item. If there are scope and content notes for items, they include the same kind of information that is in the title plus actions performed by the record. Shown below are a few of the titles of records in the Collection of Presidential Records of George H. W. Bush that have been entered into the NARA’s Archival Research Catalog.

Vice President George Bush’s Notes Regarding the Assassination Attempt on President Ronald Reagan, 03/30/1981


Letter from Eric Colton to President George H. W. Bush: 02/12/1991

Second Draft of the Address to the Nation on the Gulf War, 01/15/1991

The scope and content note of the item named by the second title is shown below.

This letter was typewritten by President George H. W. Bush and addressed to his children: George, Jeb, Neil, Marvin, and Doro. He expresses his happiness at their Christmas celebration held at Camp David, then writes concerning his conflicted feelings as he prepares for the possibility of war with Iraq.

The scope and content note includes the actions conveyed by the record, namely “expresses his happiness at their Christmas Celebration held at Camp David” and “writes concerning [expresses] his conflicted feelings as he prepares for the possibility of war with Iraq.” If the action(s) conveyed by a record could be automatically recognized, they could be used for creating the record (or item) description.

2.3 Actions and Archival Review
Under provisions of the Freedom of Information Act, Presidential records that have not been reviewed can be requested by researchers. In which case, an archivist must search the archival holdings for records that might be relevant to the request, and then review those records page-by-page for possible restrictions on disclosure. This is a time consuming and intellectually demanding task. Because of the limited human resources at the libraries there may be scores or even hundreds of pending FOIA requests.

PRA restriction a(5) “Confidential Advice” is an example of the kind of restriction on disclosure that an archivist must identify. This restriction on disclosure applies to “confidential communications requesting or submitting advice, between the President and his advisers, or between such advisers.” This includes, but is not limited to, policy or legal advice. It includes all documentary forms containing or requesting advice including final memoranda, draft memoranda, notes from meetings, letters, etc. PRA restriction a(5) applies for twelve years after the expiration of the President’s term in office.

Records that provide such advice may express the advice as recommendations, suggestions, proposals, or advice. Or they may use none of these terms at all. For example, shown below is an except from a memorandum to Chief of Staff John Sununu from Roger Porter whose subject is “Agent Orange and Vietnam Veterans” This memorandum was restricted from disclosure under PRA a(5), but has now been opened due to the expiration of the 12 year period provision of the Presidential Records Act.

The Health Policy Working Group of the Domestic Policy Council, chaired by Bill Roper, is available for monitoring the implementation of this court ruling and the new regulations. I believe there is benefit in the President publicly directing that the Working Group assume this role. This would indicate his personal interest in the issue, and that the White House is involved in the matter.

I further recommend that the President look for opportunities to speak at an appropriate event indicating his knowledge of and interest in this issue, and the fact that he has asked his Domestic Policy Council's Working Group on Health Policy to take a leadership role in this area.
This is confidential advice between advisors to the President. If the speech acts conveyed by this and similar records could be automatically identified, and they are acts such as advise, recommend, suggest or propose, then their identification would support identification of restrictions on disclosure.

In the next section, speech acts, as studied by ordinary language philosophers, logicians and linguists, will be discussed. The author hypothesizes that: The acts carried out (or performed) by a record are expressed by the record in explicit or implicit speech acts. If a method could be developed to recognize these speech acts in e-records, this would support the automatic description of records. It would also support a method for automatically identifying FOIA exemptions and PRA restrictions in e-records.

2.4 Speech Acts

Austin [5] observed that language is not only used to describe acts but to perform acts. Sentences such as the following are examples of this.

- I recommend that you attend the celebration
- I request that you attend the celebration
- I promise to attend the celebration.

Verbs like recommend, request, and promise whose action is accomplished merely by saying them or writing them are termed performative verbs.

Austin called these kinds of actions speech acts. They are to be contrasted with other human actions in which something is done as opposed to said, for example, walking, eating, gardening, etc. Among the participants in a speech act, linguists distinguish a speaker (author) who is the utterer (writer) of a message and a hearer (addressee) who is any of the immediate intended recipients of the speaker's communication.

A performative verb has a performative use in a performative sentence if the form of the verb is first person (singular or plural), present tense, indicative, active (or passive voice). For instance, in the sentence, “I promise to be there” the form of the verb is first person, singular, and present tense. An explicit performative sentence is a sentence in which the employment of a particular illocutionary force is made explicit by naming the force in the sentence itself, e.g., “I promise I shall be there.” An implicit performative sentence is a sentence in which the illocutionary force is not made explicit by naming the force in the sentence itself, e.g., “I shall be there.”

Austin claims that speech acts are of three types:

- Locutionary act: the speaking (saying) or writing of a sentence.
- Illocutionary act: an act such as recommending, requesting or promising, which is accomplished by speaking or writing a sentence.
- Perlocutionary act: an act such as persuade which affects the feelings, thoughts or actions of the addressee.

Austin also distinguished the propositional content of a message from its illocutionary force. For instance in the sentence “I recommend that you attend the celebration”, recommend designates the illocutionary force and you attend the celebration is the proposition.

Searle continued the development of a theory of illocutionary acts [6, 7]. According to Searle, every complete sentence has an illocutionary force [9 p. 7]. An illocutionary force has the seven components described below [9 p. 46].

An illocutionary point is the basic purpose of a speaker in making an utterance. According to Searle’s analysis, there are five kinds of illocutionary points:

- An assertive illocutionary point has the purpose of presenting the proposition as representing an actual state of affairs in the world.
- A commissive illocutionary point has the purpose of committing the speaker to some course of action described in the propositional content of the message.
- A directive illocutionary point has the purpose of the speaker attempting to get someone to do something described by the propositional content of the message.
- A declarative illocutionary point has the purpose of the speaker bringing into existence the state of affairs described in the propositional content of the message.
- An expressive illocutionary point has the purpose of communicating an attitude or emotion about the state of affairs described in the propositional content of the message.

The strength of illocutionary point is the strength of assertion of, commitment to bring about, direction to another to bring about, declaration of, or expression of a psychological state toward the propositional content of an illocutionary act. For example, the illocutionary force request and insist both have the directive illocutionary point, but requesting that someone do something is less strong than insisting that someone do something.

A mode of achievement is the means employed by a speaker to accomplish the illocutionary point of an utterance. For example, the means of achievement of a speaker who issues a command is his or her position of authority.

A propositional content condition is a condition imposed on the propositional content by the illocutionary force. For example, the illocutionary force of promise imposes a propositional content condition on the proposition that the proposition concerns a future state of affairs.

A preparatory condition is a state of affairs that must be presupposed by the speaker in employing a particular illocutionary force and is a necessary condition for the non-defective employment of that force. For example, the illocutionary force of a promise has the preparatory condition that the speaker presupposes his or her own capability of performing the promised action.

A sincerity condition is the psychological state of the speaker concerning the propositional content of an illocutionary act. Thus, an illocutionary act with an assertive point commits the speaker to believing the propositional content. An illocutionary act with a commissive point, e.g., promise, commits a speaker to an intention. An illocutionary act with a directive point commits a speaker to a desire. An illocutionary act with an expressive point commits the speaker to an emotional, attitudinal or sensory state.
The *degree of strength of a sincerity condition* is the strength of the psychological state that the speaker commits to in employing a particular illocutionary force. For example, a speaker who begs is expressing a stronger desire than if he requests.

An *illocutionary force indicating device (IFID)* is “Any element of natural language which can be literally used to indicate that an utterance of a sentence containing that element has a certain illocutionary force or range of illocutionary forces.” [9, p. 2]

As mentioned above Austin, distinguishes the illocutionary force of an utterance (sentence) from the propositional content. Searle and Vanderveken do as well, using \( f(p) \) to indicate the general form of simple sentences used to perform elementary speech acts, where \( f \) is the indicator of illocutionary force and \( p \) is the propositional content. For instance, in the sentence *I promise that I will come*, \( I \) promise is an IFID and \( (that) \ I \ will \ come \) is the proposition. In other words, performative verbs used in performative sentences are IFIDs.

Searle and Vanderveken use \( F(P) \) to represent the logical form of the illocutionary act itself. \( F \) is the illocutionary force and is a function of the meaning of \( f \). \( P \) is the logical form of the propositional content. \( P \) is a function of the meaning of \( p \).

Declarative, imperative and interrogative sentences are IFIDs. Statements of fact are expressed in declarative sentences; requests or commands are expressed in imperative sentences, and questions are signaled by interrogative pronouns or word order and punctuation. For example,

1. **Declarative:** You completed the report.
2. **Imperative:** Complete the report!
3. **Word order and punctuation:** Did you complete the report?

The logical form of the illocutionary act performed by sentences (1)-(3) would be:

1. **STATE** [You completed the report]
2. **REQUEST** [You complete the report]
3. **ASK** [You complete the report]

An *indirect illocutionary act* is an illocutionary act that is performed indirectly by way of performing another. For instance, a request can be made without using an imperative sentence. For example, *I need salt* or *Can you pass the salt?* in the appropriate context mean the same as the request *Pass the salt.* [7]

For textual records, as opposed to spoken utterances, textual structure can also be an IFID. For instance, in a memorandum, a section heading RECOMMENDATIONS can be an indicator that the sentences in that section have the illocutionary force recommend.

### 3. ANALYSIS OF SPEECH ACTS IN PRESIDENTIAL RECORDS

Copies of one hundred and twenty Presidential records and Personal Record Misfiles that were created during the administration of President George H. W. Bush were analyzed to determine the occurrence of

1. **explicit performative sentences,**
2. **implicit performative sentences,**
3. **speech acts indicated by textual structure,** and
4. **indirect speech acts.**

These records had already been disclosed to the public but some of them had prior restrictions on disclosure. See reference [1] for details.

Explicit performative sentences were identified by reference to a list of 271 performative verbs defined by Vanderveken [10]. Implicit performative sentences were identified by the indicative, subjunctive or imperative mood of the verb. Speech acts indicated by textual structure were identified by the occurrence of performative verbs in textual features such as section headings. Indirect speech acts were identified based on Searle’s discussion of indirect speech acts expressed in conventional forms [7]. In addition to the performative verbs defined by Vanderveken, additional verbs were identified as performative in virtue of their having an illocutionary point; that is, assertive, commissive, directive, declarative or expressive.

The analysis identified sixty-seven speech acts occurring in the records. Forty-four of these speech acts were previously identified and defined by Vanderveken [10]. These are: assert, deny, state, declare(1), tell(1), report, advise(1), remind, inform, certify(1), agree(1), acknowledge, praise(1), commit, pledge, direct, request, ask(1), ask(2), urge, encourage, invite, order(1), prohibit, suggest(2), propose, recommend, declare(2), resign, confirm, nominate, appoint, authorize, pray, terminate, veto, approve(1), disapprove, revoke, mourn, congratulate, thank, apologize, and welcome(2).

Twenty-three additional speech acts were identified and defined by the author of this paper. These are: concur, salute, amend, counsel, welcome(1), tender(2), call on, block, retire, proclaim, delegate, designate, determine, find, reject(2), endorse, appreciate, regret, trust(1), believe, want, desire, and intend.

In the following paragraphs a few of the speech acts identified in the corpus will be defined and examples of their use given. See reference [1] for additional details.

**advise(1)**

To advise the hearer that \( P \) (in its assertive use) is to assert \( P \) with the presupposition that the state of affairs \( P \) is good for the hearer. [10, p. 174]

Advise, in its assertive use, “is an official speech act, often performed (in writing) by institutions addressing individuals. It is also characteristic of lawyers, agents, and other professional people, who need to convey information to their clients and other members of the public, implying that they are acting in the addressee’s interest.” [11, p. 182]

Document 178 contains an example of a performative use of the verb advise.

This memorandum is to advise you of certain legal and policy limitations on your political activity as a member of the White House staff.

*inform*
“To inform is hearer directed in that it is to assert with the preparatory condition the hearer does not already know P.” [10, p. 175]

In document 108, President Bush informs congressional leaders of the commencement of combat operations against Iraqi forces.

Consistent with the War Powers Resolution, I now inform you that pursuant to my authority as Commander in Chief, I directed U.S. Armed Forces to commence combat operations on January 16, 1991, against Iraqi forces and military targets in Iraq and Kuwait.

certify(1)

“In the assertive sense, to certify is to assure that a proposition is true, in a formal way with the perlocutionary intention of having the hearer feel ‘certain’ of the truth of the proposition.” [10, p. 176]

Document 155, a Presidential Determination, contains a performative use of the verb certify.

Pursuant to Section 1307 of the National Defense Authorization Act, Fiscal Year 1989, I hereby certify that Saudi Arabia does not possess biological, chemical, or nuclear warheads for the intermediate-range ballistic missiles purchased from the People's Republic of China.

accept

“In general, to accept P is to commit ourselves to do P (or to permit that P be done) while presupposing (as a preparatory condition) that the hearer or some other person has requested P in previous conversation.” [10, p. 184]

In document 193, the President accepts John Frohnmayer’s resignation as Chairman of the National Endowment.

I received your letter of resignation today and, with sincere thanks and appreciation for your service, I accept your resignation effective May 1.

direct

“The verb direct names the primitive directive illocutionary force. It is generally used in the passive form as in “You are hereby directed to…” [10, p. 189] The propositional content condition is that the propositional representative of the future course of action of the hearer. The preparatory condition is that the hearer is capable of carrying out that action. The sincerity condition is that the speaker wants or desires the hearer to carry out that action.

Document 151, a National Security Review, includes an instance of the performative verb direct.

I hereby direct a short-term review of next steps for U.S. Policy towards Panama. The review should particularly focus on the implications of elections in Panama scheduled to be held on May 7, 1989. The review should include an assessment of options for the United States and recommendations for policy.

propose

“To propose that a hearer carry out some action is to suggest that he accept doing that action.” [10, p. 195]

Document 179 expresses a proposal with a first person plural possessive pronoun “Our” and the nominalized form of the verb propose.

Our proposal is to ban corporate, union, and trade association PACS, which essentially would return the law to its pre-Federal Election Campaign Act of 1971 status.

recommend

“To recommend is to advise while presupposing that the future action recommended is good in general, and not only for the hearer. When one recommends a person or thing to a hearer, one recommends that he favor that person or thing.” [10, p. 197] The direct object of recommend is usually an object, person or action.

In sample document 094, there are examples of the use of the performative verb recommend with the syntactic form “We recommend that-clause” and “We recommend gerund”.

We recommend that, in addition, they must identify a particular practice, or group of practices, that caused the disparate impact because of the race, religion, sex, or national origin of those who are underrepresented.

We recommend permitting women who have suffered on-the-job harassment to recover compensatory damages and punitive damages to the extent permitted under state law.

resign

“To resign is to renounce one’s tenure of a position, thus making it the case by declaration that it is terminated. This special propositional content condition determines the preparatory condition that one in fact occupies the position and has the power to relinquish it.” [10, p. 199]

In document 098, Sam Skinner resigns his position of Chief of Staff.

I respectfully resign the Office of Chief of Staff to the President of the United States effective August 23 1992.

proclaim

To proclaim is for somebody (with the power to do so) to declare to a wide audience a new desirable state of affairs. Vanderveken does not define the performative verb proclaim.

Document 052, a Presidential proclamation is a prime example of the use of the performative verb proclaim.

NOW, THEREFORE, I, GEORGE BUSH, President of the United States of America, by virtue of the authority vested in me by the Constitution and laws of the United States, do hereby proclaim October 1990 as Energy Awareness Month. I urge the people of the United States to observe this month with appropriate educational programs and activities.

delegate

To delegate is to declare that the authority to perform a function is transferred to another person. It presupposes that the delegator has the authority to exercise that power. Vanderveken does not define delegate as illocutionary act.
In document 191, the President delegates some authorities to the Secretary of State.

By virtue of the authority vested in me by the Constitution of the United States of America, including section 301 of title 3 of the United States Code, I hereby delegate to the Secretary of State the functions vested in me by section 303 and section 324 of the Foreign Relations Authorization Act, Fiscal Years 1992 and 1993 (Public Law 102 - 138).

4. SPEECH ACTS AND DESCRIPTION

Our hypothesis is that the speech acts identified in records support record (or item) description. To test this hypothesis, item descriptions (or scope and content notes) were manually constructed for each of the records in the corpus of 120 Presidential records. The descriptions indicate the document type, the author and addressee of the record, and the act(s) and topic(s) of the record. The action in the scope and content note was created on the basis of performative sentences, implicit speech acts, structural features of the record indicating the speech act, or indirect speech acts expressed in the records. Shown below are a few examples of the descriptions.

Signature Memorandum from Boyden Gray to the President recommending the nomination of Ronald B. Leighton to be a US District Judge.

Letter from President Bush to President Mikhail Gorbachev suggesting an informal meeting.

Memorandum from President Bush to Boyden Gray requesting an analysis of the War Powers Resolution.

Letter from Susan Black to President Bush expressing appreciation for nomination and commitment to serve.

Referral Memorandum from Sally Kelley to FEMA requesting appropriate action to a letter from Beryl Anthony to the President.

The following table summarizes the results of the test.

<table>
<thead>
<tr>
<th>Illocutionary Force Indicating Device (IFID)</th>
<th>Number of records in which the IFID was used to create the record’s description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Explicit Performative Sentence</td>
<td>77</td>
</tr>
<tr>
<td>Implicit Performative Sentence</td>
<td>31</td>
</tr>
<tr>
<td>Speech Act Indicated by Textual Structure</td>
<td>11</td>
</tr>
<tr>
<td>Indirect Speech Act</td>
<td>1</td>
</tr>
<tr>
<td>Total</td>
<td>120</td>
</tr>
</tbody>
</table>

Almost two thirds of the actions expressed in the scope and content notes could be determined from explicit performative sentences. About one fourth of the actions could be determined from implicit performative sentences. About one tenth of the actions could be determined from speech acts indicated by structural features. Only one act in the scope and content notes required the recognition of an indirect speech act. All actions expressed in the scope and content notes could be determined from the speech acts identified in the records.

One of the research issues that arises is: Since every complete sentence expresses a speech act, how does one choose the appropriate act(s) for inclusion in the scope and content note? It is our current hypothesis that the solution lies in use of the strength of illocutionary point of the speech act. Recall that this is the strength of assertion, commitment to bring about, direction to another to bring about, declaration of, or expression of a psychological state toward the propositional content of an illocutionary act. For example, the illocutionary force request and demand both have the directive illocutionary point, but requesting that someone do something is less strong than demanding that someone do something. The concept of degree of strength will need to be extended across illocutionary points. For instance, to represent that a directive or declarative is stronger than an assertive, commissive or expressive illocutionary point.

5. A METHOD FOR RECOGNIZING SPEECH ACTS IN E-RECORDS

Our goal is develop a method that will annotate e-records with a representation of the speech acts (illocutionary force and proposition) performed by sentences in e-records. This representation will be used with annotations of the topics of sentences, paragraphs and sections of e-records to develop a method for automated item description [3]. The representations of speech acts and topics of sentences in e-records will also be used in a prototype system for automatic reasoning to determine possible FOIA exemptions and PRA restrictions on disclosure of the contents of Presidential e-records [12].

The analysis summarized in section 3 indicates the kinds of speech acts that occur in Presidential records. The performative sentences and IFIDs identified provide criteria for recognizing these speech acts.

The method for recognizing speech acts in e-records is based on a method developed in prior and ongoing research to identify the documentary form (genre) of e-records and thereby extract descriptive metadata such as author’s and addressee’s names, topic and date [2]. The additional methods for recognizing speech acts consist of the following processing resources.

- **Pronominal Coreferencer**
  - Supple Parser + grammar for English + interpretation rules

- **Speech Act Transducer**

Pronominal coreference resolution begins by resolving pronouns such as I, we, my, our, you and your to the names of the author(s) and addressees of the record.

Then a parser with a grammar for English is used to produce a representation of the syntactic form of sentences in a record and a quasi-logical representation of the semantics of those sentences.

The function of the speech act transducer is to produce a representation of the speech acts performed by the sentences in the record and from these to create a representation of the primary speech act(s) performed by the record as a whole.

The transduction of the syntactic structure of a sentence and its quasi-logical form into an annotation of the speech act of that
sentence is accomplished by the Java Annotation Pattern Engine (JAPE) and JAPE rules. The JAPE rules are processed by the Java Annotation Pattern Engine in phases. The rules in the first phase are processed before the rules in the second phase, and so on. Six phases are anticipated.

Verb Sense Disambiguation
Annotation of Explicit Speech Acts
Annotation of Implicit Speech Acts
Annotation of Speech Acts Indicated by Text Structure
Annotation of Indirect Speech Acts
Annotation of the Primary Speech Acts Performed by the Record

The speech act transducer must first determine whether the verb is ambiguous, and if so disambiguate it. For instance, the verb agree has three related but different meanings as a performative verb. First, a person can be in agreement with something somebody else said. For example, “I agree with Senator Cook that the resources of the nation belong to the nation, not to the multinational oil companies or to any individual.” Second, a person can agree to do something or agree to a condition. For example, “I agree to attend the meeting.” Third, persons with different ideas as to how to do something can by mutual concession or discussion agree on the same solution. For example, “We agree on whom to elect chairman.” The sentence pattern “I (or we) agree that CLAUSE” can be used to recognize the second meaning. The sentence pattern “I (or we) agree to INFINITIVE” can be used to recognize the first meaning. The sentence pattern “I (we) agree PREPOSITIONAL PHRASE” can be used to recognize the third meaning. The form against which to match the patterns is provided by SUPPLE as a parse tree of a sentence.

The simplest type of speech acts to recognize is one in which the speaker (writer) explicitly uses a performative verb in a performative sentence. Most sentences in the first person present indicative containing a performative verb are performative sentences, e.g.,

I promise to come on Wednesday.

One also must recognize the nominalized forms of performative verbs.

My declaration is that

It is our statement that

My question is

There are also some occurrences in the present continuous, e.g.,

I am asking you to do this for me.

Some performative sentences use the verbs in the plural

We pledge our lives, our fortunes and our sacred honor.

Furthermore, some performative sentences are in passive voice, e.g.,

Passengers are hereby advised that all flights to Phoenix have been cancelled

Sometimes the performative expression is not a verb and it may be in a separate clause or sentences, as in

I’ll come to see you next week and that is a promise

The next simplest speech act recognition task is the recognition of implicit speech acts that take the form of declarative, imperative and interrogative sentences. For example,

1) Declaratives that are not performative sentences are usually assertions.

2) An imperative is always a request or command, e.g., "Let’s discuss this." or "Let me know what you think."

3) Interrogatives are usually requests for information, or equivalently, asking a question.

Speech acts are sometimes represented by text structure rather than in sentences. For instance, Document 145 in the corpus is an example in which the speech act of recommending is indicated by a run-in paragraph heading, and the use of the auxiliary verb should.

RECOMMENDATION: We should submit, or at least enthusiastically endorse, legislation that subjects Congress to the same laws it has imposed on the rest of the people. We should also make greater mention of our support for term limitations.

The procedure for document type recognition, which is performed prior to the speech act transduction process, produces annotations of a record representing its textual structure. The next phase of speech act recognition examines that textual structure to determine whether it is indicating speech acts.

The next phase is to recognize indirect speech acts. Indirect speech acts are commonly used to make requests and to reject proposals. Rules can be formulated for recognizing when an interrogative or declarative sentence is a request or command. For example

1) An interrogative containing a modal auxiliary "can," "could," "will" or "would," with the addressee as subject and the predicate describing an action that is feasible for the addressee to perform is to be interpreted as a request or command.

2) A declarative in which the subject is the speaker (author), e.g., "I," or the speaker and the addressee(s), e.g., "We," and the verb is "need," or "want," is to be considered a request or command.

Additional examples will be sought of the occurrence of indirect speech acts in Presidential records. These will be used to refine the indirect speech act recognition rules.

The prior phases of the speech act transducer produce annotations of the speech acts performed by individual sentences and sequences of sentences in the record. The last phase contains JAPE rules that transform these annotations into a logical representation of the primary speech act(s) performed by the entire record. This is combined with metadata extracted during document type recognition. The result is of the form:

[document(e1), author(e1, S), addressee(e1, H),
act(e1 F(P)), topic(e1, T), date(e1, D)]

This notation means that e1 is a document, the author of e1 is S, the addressee of e1 is H, the primary action of e1 is F(P), the topic of e1 is T, and the date of e1 is D. This representation will be
passed to software modules for automatic record description and checking for access restrictions.

6. SUMMARY
This research has answered the question: How are actions conveyed by records? The answer is: Through the speech acts expressed in the records. It has also shown that the speech acts identified in records can be used with other elements of the records to create record descriptions. It has raised the research issue: How does one select the primary speech acts in a record from all the speech acts in the record? It is suggested for further investigation that the answer lies in the strength of the illocutionary points of the speech acts.

The implication of these findings for archival and digital curation curricula is significant. An understanding of the nature of records is enhanced by concepts from the theory of speech acts. Language is used to perform acts. Every complete sentence is a speech act. The purposes of speech acts are to make assertions, commitments, declarations, give directives or express feelings and attitudes. The description of records (or aggregations of records) in the form of scope and content notes is based on recognizing the primary speech acts of the records.

The development of techniques for recognizing explicit and implicit performative sentences and indirect speech acts are important research issues in the field of Computational Linguistics. So also are techniques for automated summarization of textual documents.

Intellectual control of an archival collection is not achieved until the collection has been fully described. The practical significance of a speech act recognition technique lies in the potential for automated record description and thus in earlier intellectual control of archival collections.

Archival review decisions are based on knowledge of the identity of the author(s) and addressee(s) of a record, the context of a record, and the actions and topics expressed in a record. The automatic recognition of speech acts in records is an enabling technology for decision support of archival review.

The method described in this paper for automatically recognizing the speech acts expressed in e-records will be implemented, tested and experimentally evaluated on records selected from a collection of Presidential e-records. Assuming that it achieves a high level of performance, the implemented method will be used in further development of methods for archival description and review.

7. ACKNOWLEDGMENTS
The Army Research Laboratory (ARL) and the National Archives and Records Administration (NARA) sponsor this research under Army Research Office Cooperative Agreement W911NF-06-2-0050. The findings in this paper should not be construed as an official ARL or NARA position unless so indicated by other authorized documentation.

8. REFERENCES
Reconstructing the Digital Past: A Case Study of the Reconstruction of the Lost Pittsburgh Project

Bernadette G. Callery
School of Information Sciences
University of Pittsburgh,
Pittsburgh, PA 15260 USA
bcallery@sis.pitt.edu

ABSTRACT
The Web-based Pittsburgh Project, aka Functional Requirements for Evidence in Recordkeeping, was administered by the University of Pittsburgh’s School of Information Sciences between 1992 and 1996. The site disappeared in 2000 when the School switched servers. Although partial versions of the project could be recovered through the Wayback Machine, graduate students in the School’s 2008 Digital Preservation course reconstructed the entire site, and added documentation on the reconstruction process. The reconstructed site is now available at http://www.sis.pitt.edu/~bcallery/pgh/index.htm. This case study discusses educational strategies used in the reconstruction process, particularly the introduction of issues of the completeness and authenticity of the restored site, and considers the effectiveness of collaborative tools in the management of a group project.

Categories and Subject Descriptors
K.3.2. Computer and Information Science Education. K.4.3. Organizational Impacts

General Terms
Management, Documentation

Keywords
Pittsburgh Project, Functional Requirements

1. INTRODUCTION
One of the major concerns with the maintenance of digital documents is the ease with which they can be deleted or altered. While some web-based records can be recovered through the use of salvage sites such as the Wayback Machine, users should then question whether the files so retrieved actually represent the final version intended by the authors. In collaborative sites, these questions become more complex, as users are then dependent upon the availability of multiple inter-related sites.

This case study examines the reconstruction of the Web-based Pittsburgh Project, or Functional Requirements for Evidence in Recordkeeping, funded by the National Historical Publications and Records Commission (NHPRC) between 1992 and 1996, led by Richard J. Cox, University of Pittsburgh, School of Information Sciences. The purpose of the original Pittsburgh Project was to investigate possible techniques for determining the authenticity of electronic documents. Specific outcomes of the Project’s research included functional requirements for recordkeeping systems, production rules, metadata specifications and the concept of a literary warrant which reflected the “professional and societal endorsement of the concept of the recordkeeping functional requirements” [1].

The resultant Pittsburgh Project website brought together the working papers, progress reports, a bibliography and a list of project participants. The original site included only citations to external publications related to the project and produced by the participants, but not the text of those publications. The project concluded in 1996 after which no further updates were made to the website. The loss of the entire site occurred in 2000 when the project files were not transferred when the School switched servers. Prospective users searching for the site were referred to the Internet Archive’s Wayback Machine in order to locate a version of the site as captured in one of its Internet crawls.

2. RECOVERY PLAN
Rather than simply rebuild the lost site, the students were advised that the restored original site should appear as the core of a new website, surrounded by documentation about the process of reconstruction. This design decision illustrated the need to distinguish the recovered parts of the original site from the additions of updated links and new material. As the recovery team noted in their documentation, “We felt that it was more authentic to note what was there, rather than to force an inauthentic recovery.”

The initial efforts of recovery of this site included searching the Internet Archive Wayback Machine for archived copies of the site. Versions of the site were captured in eleven separate crawls of 1998, 1999 and 2000, with the version captured in 1998 identified as being the most complete [2]. Difficulties with the intermittent access to and availability of those Wayback Machine sites reinforced the need for a stabilized reconstruction of the Pittsburgh Project site. Additional searches of the print and online literature retrieved external references to the project, its participants and its reception in the archives and
information science community. This updated list of references was incorporated into the restored site.

2.1 Editorial Decisions
As noted in the restoration project documentation, “The most challenging task was identifying what, exactly, we were recovering. Unlike traditional paper-based documents, websites change over time, are non-linear and contain external links” [3]. Although the 1998 Wayback Machine capture of the site was reasonably complete, it contained broken links and lacked the lists of functional requirements. To distinguish between which documents were present and which were represented only by broken links, a key document was created which listed the titles of the website’s documents in a hierarchical list which in turn reflected the structure of the original website. Missing items were indicated by a strike-through font. External links which remained unlocated at the conclusion of the project were noted as inactive and irretrievable in this key to indicate the need for future work.

Where broken links were included in the restored website, customized error messages indicated the current URL address if the content had been located elsewhere. Unresolved broken links noted the unsuccessful attempt to find the information and solicited the user’s assistance in locating the missing content.

2.2 Workflow Management
Strategies for successful project management evolved over the course of the project. The major challenge was to organize working groups that were sufficiently separate so that their work did not overlap. One successful approach was the formation of Inside and Outside Committees. The Inside Committee, which dealt with contributions by University of Pittsburgh faculty and technical staff, met with current representatives of those groups and searched for publications related to the project so that they could be linked to the reconstructed site. The Outside Committee identified external contributors (i.e. non-Pitt faculty), and sought their permission to re-publish their work in this new context. As there had been no special agreements made for the transfer of copyright, it was presumed that copyright continued to reside with the original authors, or, in the case of authors associated with the governmental agencies, with those agencies.

2.3 Collaborative Tools
The online course management and delivery tool used for this course was Blackboard, whose version 7.3 includes a number of Web 2.0 collaborative tools. A project wiki was built to maintain documentation on the project and to co-ordinate the efforts of the various members of the project team. The wiki was used to collect the reports of the various committees, meeting minutes and versions of documents under construction. Other sections of the wiki maintained the ongoing task list, comments from the instructor and a status report as required by the assignment.

While the project team attempted to use the wiki as a means of coordinating their activities, reporting progress, and sharing and editing documents, it proved difficult to use as a current awareness tool. The principal objection was that it was difficult to determine what pages or other content had been newly added. In time, this information was communicated to the group members via email, as too much time was spent in the project meetings making sure that all the participants were up to date with the content on the website.

2.4 Comments from the Original Authors
Several of the original contributors who responded to requests for copies of their documents from the original project noted that they had published elsewhere on the project and therefore there was little need to reconstruct the site. In reporting their interview with project director Cox, the Inside Committee noted that “he feels that anything useful about the project was written in the essays...He has little nostalgia about the project and mentioned that when it was lost, his attitude was that the project had ended anyway.”

2.5 Distinguishing the Original
As in conservation of museum objects, it the restoration must be clearly distinguished from the original object. The original website, embedded in the restoration site, retained its 1996 look and feel. The 2008 wrapper website is quite distinct in its overall design, with the dark blue background providing a strong contrast to the simpler original site. The home page of the restoration provided background on the significance of the original project, and gave an overview of the restoration process, including the addition of specific information to complete the broken links.

3. CONCLUSIONS
In this project, students acknowledged that the ideal of a complete restoration was not achievable and focused on stabilizing what could be located and identifying supplementary information. Comments from the site’s original creators indicated that not all agreed that a permanent record of this manifestation of the Pittsburgh Project was necessary or desirable. These responses from the creators raise the question of whether or not all lost digital projects should be found, reminding us that appraisal decisions continue to apply in the digital realm.

4. ACKNOWLEDGMENTS
My thanks to the students in the 2008 Digital Preservation course at the University of Pittsburgh, School of Information Sciences who worked on this project: Jessica Benner, Cathy Cleveenger, Sara Deegan, Ross Griffiths, Sarah Hackett, Brandi Liskey, Jenny Lockerby, Erin Mackin, Abigail Middleton, Miranda Nixon, Ann Retzinger, Erin Thompson.

5. REFERENCES
Identifying and Implementing Modular Repository Services

Leslie Johnston
Library of Congress Office of Strategic Initiatives
Washington D.C., 20540 USA
lesliej@loc.gov

ABSTRACT
In recent work at the Library of Congress, we have been identifying requirements for digital repositories for locally created collections and collections received from partner institutions. Our most basic needs are not surprising: How do we know what we have, where it is, and who it belongs to? How do we get files – new and legacy – from where they are to where they need to be? And how do we record and track events in the life cycle of our files? This paper describes current work at the Library in implementing tools to meet these needs as a set of modular services -- Transfer, Transport, and Inventory -- that will fit into a larger scheme of repository services to be developed.

1. INTRODUCTION
In examining the reasons why more institutions do not have trusted digital repositories for their collections, one always ends up with a long list of daunting requirements for building such a beast. Where do you begin? What are your most basic needs? What is the first step? In recent work at the Library of Congress, we have determined that our most basic needs are really basic: How do we know what we have, where it is, and who it belongs to? And how do we get files – new and legacy – from where they are to where they need to be?

The Library of Congress Office of Strategic Initiatives has been working on solutions for a category of activities that we refer to as “Transfer.” At a high level, we define transfer as including the following human- and machine-performed tasks:

- Adding digital content to the collections, whether from an external partner or created at LC;
- Moving digital content between storage systems (external and internal);
- Review of digital files for fixity, quality and/or authoritativeness; and
- Inventorying and recording transfer life cycle events for digital files.

Transfer processes are not surprisingly linked with preservation, as the tasks performed during the transfer of files must follow a documented workflow and be recorded in order to mitigate preservation risks. Defining, implementing, and documenting appropriate transfer processes depends on the requirements of each collection building project, which can vary wildly. Best practices are still emerging.

2. WHERE IS THE CONTENT COMING FROM?
The Library of Congress has been digitizing its collections for over 15 years -- making collection materials available online since 1994 starting with the “American Memory” site\(^1\) -- concentrating on its most rare collections and those unavailable anywhere else [1]. The collections include photographs, manuscripts, maps, sound recordings, motion pictures, and books, as well as "born digital" materials.

The Library of Congress is involved in a number of collection-building activities with external partners. The National Digital Information Infrastructure and Preservation Program (NDIIPP)\(^2\) provides funding through the National Science Foundation to over 130 institutional partners that send content to LC for stewardship and preservation [1]. Content under stewardship by NDIIPP partners includes geospatial information, web sites, audio visual productions, images and text, and materials related to critical public policy issues. Each partner may deliver its content by a different transport mechanism, e.g., shipping hard drives or network transfer. Once acquired, the digital content must be validated and verified, inventoried, and placed in archival storage.

The National Digital Newspaper Program (NDNP)\(^3\) is a partnership between the NEH and the Library of Congress to provide enhanced access to United States newspapers [2]. Over a period of approximately 20 years, NDNP will create a national, digital resource of newspapers from all the states and U.S. territories published between 1836 and 1922. In the NDNP transfer process, packages of digitized newspaper prepared by awardees are delivered on hard drives.

In 2004, the Library’s Office of Strategic Initiatives created a Web Capture team\(^4\) to support the goal of managing and sustaining at-risk digital content. The team is charged with building a Library-wide understanding and technical infrastructure for capturing Web content. The team is identifying policy issues, establishing best practices and building tools to crawl, collect, and preserve Web content. As of fall 2008 the team has completed 17 Web archive collections and is working on building Web archives for four collections comprising approximately 3000 Web sites.

\(^1\) The American Memory collections are available at: http://memory.loc.gov/ammem/index.html.
\(^2\) For information on NDIIPP, please see: http://www.digitalpreservation.gov/.
\(^3\) For information on NDNP, see: http://www.loc.gov/ndnp/.
\(^4\) For information on the Library’s web capture activities, see: http://www.loc.gov/webcapture/.

This work is licensed under the Creative Commons Attribution-Noncommercial-No Derivative Works 3.0 Unported license. You are free to share this work (copy, distribute and transmit) under the following conditions: attribution, non-commercial, and no derivative works. To view a copy of this license, visit http://creativecommons.org/licenses/by-nc-nd/3.0/.

DigCCurr2009, April 1-3, 2009, Chapel Hill, NC, USA
3. DEPOSIT AND TRANSFER TOOLS

Content transfers to the Library have largely consisted of small numbers of bulk transfers of content – typically, tens to hundreds of gigabytes at a time – using varied, primarily manual processes driven by Library staff to pull the content into the Library’s environment. Through these experiences we have gained experience in transfer processes, and have been able to simplify the variety and complexity of the initial approaches to such bulk transfers. As the content packaging, transfer, and inventorying approaches have standardized, it has become possible to plan to grow and automate the number of transfers, in order to support new transfer scenarios in which transfers might involve many pushed deposits at smaller scale than the current bulk transfers.

For the purpose of transferring content to the Library, a package is a set of files stored in a file system, which may be a subset of a larger collection of content, to be transferred and managed as a unit. The set of files comprising a package may be transferred as a single file in a container format such as ZIP or tar to be unpacked upon receipt. Working with John Kunze of the California Digital Library, Andy Boyko, Justin Littman, Liz Madden, and Brian Vargas of the Library produced a generalized version of what had been initially referred to as the “LC Package Specification,” now called “BagIt” [3-4].

The base directory of a Bag contains a file manifest, a content directory, and an optional package information directory. The content directory contains the contents of the package, as defined by its producer. The content directory may have any name and internal structure. There is no limit on the number of files or directories this directory may contain, but its size should make practical transfers easier, based on physical media limitations or expected network transfer rates. In the Library’s experience, 500 Gb is the recommended maximum size, although Bags as large as 1.8 Tb have been transferred.

The file manifest lists the names and checksums of the content files and the package information files, excluding itself and any shipping files. The Library has been working primarily with md5deep, but any commonly recognized cryptographic checksum algorithm can be used to generate the manifest. Neither the file manifest nor the package manifest obviates the need for descriptive metadata being supplied by the package producer. The manifests assist in the transfer and archiving of the package as a unit, rather than supplying any description of the content.

Bags may be created before or after the act of transfer. The creation and communication of original checksums that accompany the Bag for verification after completion of the file transfer support a more easily audited process. In-house, content that is not received in Bags will have Bags generated as an aid to verify internal transport to archival systems and to aid in file-level preservation.

The Library has created a Bag Validator script, which checks that all files listed in manifest are in the data directory; there are no files in the data directory that are not listed in manifest; and there are no duplicate entries in the manifest. The VerifyIt script is used to verify the checksums of files in a Bag against its manifest. The Bag Validator is a Python script and VerifyIt is a shell script.

A client-side Bagger application has been developed to assist partners engaged in small-scale deposit transfers, automating the packaging and submission of locally-hosted content without requiring Library involvement, and ideally requiring no client-side IT support or infrastructure. This tool will be equally suited for packaging and transferring internal LC content, such as DVD or CD archives, to centralized transfer and storage environments. It is implemented as a Java Web-Start application for use across platforms, and supports the aggregation of files into Bag packages, including the creation of checksum manifests and Bag information files. This application was in part built on top of BIL—the BagIt Library—a Java library developed to support Bag services.

In order to support the expanding numbers and types of transfers, several software tools were needed to help automate transfers. The Deposit service is a web-hosted application for use by transfer partners in registering a new transfer; this application will support the registration and initiation of the transfer content via network transfer (rsync, ftp), and via fixed media, such as hard drives or DVDs. The web application is implemented using the Django web framework. At the time of this writing, Deposit services are mid-way through the production implementation process, including review by representatives of the multiple digital content acquisition projects.

The Deposit tools are tied to a series of Transfer and Transport back-end tools used for retrieval, receiving and managing of content transfers. The Parallel Retriever implements a simple Python-based wrapper around wget and rsync, capturing files and producing a package that meets the BagIt specification when given a "file manifest" and a "fetch.txt" file. It has been used to transfer content from several transfer partners hosting rsync and HTTP servers, at rates exceeding 200Mbps over Internet2. It was initially built specifically for transfers from the Internet Archive to the Library via rsync, but has been extended to HTTP and FTP.

Underlying “Core Transfer” components support various transfer functions. The components are completely independent of any workflow, though they may of course be invoked by any designated workflow. “Core Transfer” services provide a container for running transfer components so that they can be invoked though a Java Remote Service and respond with the Service Request Broker. It is implemented using Spring and Hibernate.

4. INVENTORY TOOLS

The goal in developing Inventory Tools is to satisfy needs identified through the process of doing transfers manually and attempting to record their outcomes. These include keeping track of package transfers for a project, tracking individual packages and events associated with them, and a list of the files that make up each package and their locations. For legacy collections these tools can be pointed at existing directories to package, checksum, and record inventory events to bring the files under initial control.

The Inventory System has three parts: the Package Modeler, a suite of command line inventory tools, and a reporting web
application. The Package Modeler implements a domain model for packages (Projects, Packages, Canonical Files, File Locations, and File Instances) which can be recorded in a persistent way, updated, and queried. The Package Modeler is implemented using Java objects mapped to a PostgreSQL database using Hibernate for object-relational mapping. The Inventory tools inspect packages and update the Package Modeler. The command line tools are called by the Package Modeler. The reporting web application allows users to view reports on packages.

Since the Package Modeler must also represent the history of a package, it must record events. There are events that occur on a Package level (Package Events) and on a file level (File Location Events). Examples of Package Events include “Package Received Events,” which are recorded when a project receives a package; and “Package Accepted Events,” which are recorded when a project accepts curatorial responsibility for a package. Examples of File Location Events include “File Copy Events,” which are recorded when a package is copied from one File Location to another; and “Quality Review Events,” which are recorded when quality review is performed.

The Transfer, Transport, and Inventory tools can be tied together into any of a number of project-based Workflow systems. The underlying workflow engine is jBPM\textsuperscript{10}, an open-source workflow system. The drivers of a workflow are process definitions, which represent the process steps. jBPM Process Definition Language (jPDL), the native process definition language of jBPM, is used to encode the workflow process steps as XML. The workflow can be designed using the visual editor Graphical Process Designer, a plug-in for the Eclipse platform.\textsuperscript{11} A web user interface, called the Transfer UI in its first implementation for the NDNP, allows users to identify lists of tasks to be performed, initiate, monitor and administer processes; and notify the workflow engine of the outcome of manual tasks, including task completion. Workflow tasks instantiated through the system include transfer, validation by an NDNP-specific validation application \textsuperscript{5}, manual quality review inspection, and file copying to archival storage and production storage. The Transfer UI was implemented using Spring MVC. Both the Inventory system and all workflows are built on top of the BIL Java Library.

5. CONCLUSIONS
At the time of this writing, the Transfer, Transport, Inventory and Workflow services have been put into production for NDNP \textsuperscript{5}. The Transfer and Transport tools have been put into production for NDIIPP, and production implementation is under way for the Inventory, Deposit, and Workflow services, as well as the Bagger application. By the end of 2008, many incoming collections will be processed using these tools. The expectation is that a retrospective inventory of the Library’s digital collections will be undertaken in 2009 using these tools.

The first of the Transfer tools—the Parallel Retriever, the Bag Validator, and VerifyIt—have been released by the Library of Congress as open source on SourceForge\textsuperscript{12}. Additional tools and utilities will be released over time.

Why are such transfer tools and processes so important? While our initial interest in this problem space came from the need to better manage transfers from external partners to the Library, the transfer and transport of files within the organization for the purpose of archiving, transformation, and delivery is an increasingly large part of daily operations. The digitization of an item can create one or hundreds of files, each of which might have many derivative versions, and which might reside in multiple locations simultaneously to serve different purposes. Developing tools to manage such transfer tasks reduces the number of tasks performed and tracked by humans, and automatically provides for the validation and verification of files with each transfer event.

Why are we looking at close integration between transfer and inventory functions? Inventorying and audit functions have been identified as a vital aspect of data curation. One example initiative is the JISC Data Audit Framework project\textsuperscript{7} \textsuperscript{2}, where work is proceeding on the development of a registry component intended for recording the results of data audits based on the framework, which will provide organizations “with the means to identify, locate and assess the current management of their research digital assets.”

Inventory services can bring several benefits, including collection risk assessment and storage infrastructure audits. Realizing any benefits for effective data management relies on knowledge of data holdings. Knowledge of file-level holdings and recording of life cycle events related to those files from the moment that they enter the collection and in every future action reduces future risk by storing information that can be used in discovery, assessment, and recovery if and when a failure occurs.

Identifying needed services as modular rather than monolithic has allowed the Library of Congress to research and implement each of these functions in a more nimble way, all the while planning to fit those services into a larger scheme of repository services. The integration of modular transfer and inventory services as well as workflows allows for separation of tasks based on project or collection or format needs while supporting backend data integration where required. Modules can be independently reimplemented in the future when the need arises. This also allows for extensions to services and functionality that we have not yet considered, let alone planned for.

But do these services make up a repository? Looking at the OAIS Reference Model \textsuperscript{6} and Trustworthy Repositories Audit & Certification: Criteria and Checklist \textsuperscript{7}, we can consider the sections of the “Audit Checklist” at a very high level. Section A covers issues of administrative responsibility, organizational viability, financial sustainability, and procedural accountability. Section C covers criteria for a secure and trusted infrastructure. Section B covers the digital object management responsibilities of a repository. It is in this area that the modular work that the Library is undertaking in Deposit, Transfer, Transport, and Inventory can be categorized.

These modular services do not equate to everything needed to call a system a repository. There are only detached end-user discovery and delivery applications. Descriptive metadata is not yet tracked with the media files. There are currently no granular rights and access policies nor means to enforce them. Preservation

\textsuperscript{10} http://www.jboss.com/products/jbpm.
\textsuperscript{11} http://www.eclipse.org/.
\textsuperscript{12} http://sourceforge.net/projects/loc-xferutils/
monitoring is not yet in place. But there is a set of services that
equate to many aspects of “ingest” and “archiving” – the registry
of a deposit activity, the controlled transfer and transport of files,
and an inventory system that can be used to track files, record
events in those files’ life cycles, and provide basic file-level
discovery and auditing. Through the Inventory tools we expect to
be able to provide persistent access at a file level. In other words,
it may not yet be a full-blown repository, but is the first stage in
the development of a suite of tools to help the Library ensure long-
term stewardship of its digital assets.

6. REFERENCES


Web Access for the Museum of Anthropology’s Collections

Stephen L. Whittington  
Museum of Anthropology  
Wake Forest University  
Winston-Salem, NC 27109  
USA  
whittisl@wfu.edu

Kyle Elizabeth Bryner  
Museum of Anthropology  
Wake Forest University  
Winston-Salem, NC 27109  
USA  
brynerke@wfu.edu

Beverlye H. Hancock  
Museum of Anthropology  
Wake Forest University  
Winston-Salem, NC 27109  
USA  
hancocbh@wfu.edu

M. Robert Vidrine  
Instructional Technology  
Wake Forest University  
Winston-Salem, NC 27109  
USA  
vidrinmr@wfu.edu

ABSTRACT
This paper is a case study of a phased project to digitally curate archaeological and ethnographic collections and associated archival materials of the Museum of Anthropology at Wake Forest University and make them freely accessible through the World Wide Web. Multiple federal grants and infrastructure support from the museum’s parent organization have been essential for successful project implementation.

Categories and Subject Descriptors
H.2.4 [Database Management]: Systems – object-oriented databases;  
H.2.8 [Database Management]: Database Applications – image databases;  
H.3.5 [Information Storage and Retrieval]: Online Information Services – Web-based services;  

Keywords
Museum, archaeology, ethnography, archives, re:discovery.

1. INTRODUCTION
Museums are confronted with the need and opportunity to present their collections in new and innovative ways to the public. One means of developing new audiences and presenting more information to current audiences is to develop Web content and digital resources. Web content and digitization performs multiple functions for museums. Digitization provides an easier means to organize, sort, and search information about collections, but also provides a means to preserve information (Drewes, ed. 1997). Utilization of custom-made database management programs allows museums to input and manipulate large amounts of data, including text and digital images, in an efficient and easy manner. The ability to upload and digitize important information about objects in one comprehensive database assists museums in their important custodial function of preserving collections (Gorman, ed. 2007). The process of digitizing information about collections creates digital assets that must be cared for in a similar manner to the actual three-dimensional objects they document (Buck, ed. 1998). Therefore, digitization becomes a standardized and systematic means for a museum to document existing collections, preserve that information, and create new information about collections in a digital format. The following is a case study of how one museum developed and implemented a database system and online information service to promote and preserve its collections.

2. MUSEUM OF ANTHROPOLOGY
The Museum of Anthropology at Wake Forest University creates awareness of global cultures by collecting, protecting, managing, and exhibiting archaeological artifacts, ethnographic objects, and visual arts of past and present peoples, and providing opportunities for intercultural learning.

Founded in 1963, the Museum of Anthropology (MOA) was established by the faculty of the Department of Anthropology at Wake Forest University to broaden learning opportunities for students. Since its inception, the MOA’s educational role has expanded greatly within the Winston-Salem and Piedmont Triad communities, to the extent that we are “Winston-Salem's global cultures center.” In its present location on campus since 1986, the MOA's collections, exhibits, and programs have grown extensively. Long-term exhibits currently consist of objects from North America, Asia, and Oceania. Two to three times a year the special exhibits gallery houses topical exhibitions reflecting university and community interests. The staff of the MOA offers the following educational programs: in-house and outreach curriculum-based programs for grades K through 12 during the school year; Sunday afternoon classes for families during the fall and spring; week-long thematic camps for children in grades 1 through 5 during the summer; and portable outreach boxes for use in the classroom or at cultural festivals. In addition, the MOA offers a variety of scholarly and popular lectures, workshops, and family and Scout days each year.

During the July 1, 2007–June 30, 2008 fiscal year, the staff of the MOA served 1,158 Wake Forest students, 4,688 K-12 students from Forsyth and surrounding counties, 789 K-12 adult chaperones, and 3,607 others (adults, seniors) in the museum, and 2,755 K-12 students, 254 K-12 adult chaperones, and 8,154 others off-site. In addition, the MOA indirectly served 10,221 people...
through outreach boxes and 17,100 Wake Forest students, faculty, and staff through exhibits elsewhere on campus. As an estimate of the diversity of the MOA’s audience, for K-12 children coming to the museum 54% were Caucasian, 26% were African American, 16% were Hispanic, 3% were Asian American, and less than 1% each were Native American and disabled.

### 2.1 Collections

The MOA’s collections of some 28,000 archaeological and ethnographic objects represent ancient and contemporary traditional cultures from around the world. No other museum in North Carolina collects from such a diversity of world areas, past and present. The first and only exposure many people have to these cultures is through exhibits and outreach based on the museum’s collections of stone tools, wooden masks, woven textiles, metal adornments, and ceramic vessels. Subsets of the collections reveal important facets not available elsewhere of the history of the area’s Moravian community and the early development of archaeology in North Carolina. The collections are used for teaching university and K-12 students, in public outreach, for long-term exhibits and for loans to other institutions, and are the basis of scholarly publications and academic theses.

#### 2.1.1 Archaeological Objects

The archaeological collection primarily consists of 20,517 artifacts collected in the 1930s and 1940s by Douglas Rights, a founder of the North Carolina Archaeological Society and a Moravian minister. Artifacts in his collection come primarily from the Western Piedmont and Yadkin-Pee Dee River basin of North Carolina and represent the Archaic and Woodland peoples of this area. There are 1,150 objects from the Americas south of the United States. We have 17 Bura (3rd-11th centuries) ceramic reliquary objects from Niger, one of the only collections of Bura material in the United States. In addition, there are 1,188 objects from Europe, North America, and unknown locations.

#### 2.1.2 Ethnographic Objects

The ethnographic collection of 5,106 objects represents cultures located throughout most of the non-Western world. Museum staff members, Department of Anthropology faculty and students, and private individuals collected some objects within their cultures of origin and documented their significance. The largest single ethnographic collection is from Africa, consisting of 1,474 objects. Significant portions of the 1,121 objects from Mexico, Central America, and South America consist of Maya textiles from the 1970s to the present, objects used in Día de los Muertos celebrations in Mexico, and Amazonian personal adornment from Brazil. The collection includes 1,239 objects from Asia: 654 late Q’ing dynasty items from China, 180 objects from Japan, and the remainder from scattered areas of Southeast Asia. There are 426 ethnographic objects from North America, including Yup’ik and Inuit materials collected by Moravian missionaries in Alaska and Labrador. Oceanic artifacts made by many island peoples, particularly those of New Guinea, number 161. The remaining items are from Australia, northern Pakistan, India, the Middle East, and unknown areas.

#### 2.1.3 Archives

The archival collection consists of almost 7,000 paper documents, printed photographs, slides, and negatives that relate to and support the object collections. Generally the materials in the archives were created by the collectors of the objects and record information that makes the function and significance of the objects more comprehensible to people not familiar with their cultures. As with the objects, many of the archival documents and images record information about life-ways that are no longer practiced and cultures that are changing rapidly under the effects of globalization. The archives contain irreplaceable information about cultural and environmental contexts of our collections of objects from past, changing, and modernizing cultures. It is critical for the documents and photographs to be carefully maintained, and the data they contain accurately and completely preserved and made accessible to people living in communities around the museum, many of which include representatives of those cultures.

### 3. Development of the Digital Database

Until 2005, the MOA had no Registrar or Collections Manager and creation of paper and computer database records fell to staff and students when they could spare time from their regular duties. Distractions and inadequate training led to many errors and inconsistencies in both physical and digital records. Few objects had been photographed. In addition, because of a history of trying several data management programs, all developed primarily for business applications and designed in-house for financial reasons, digital records did not have consistent lexicon categorization and some data from physical records were not in the computer catalog because of insufficient field space.

Eventually, the museum settled on ACCESS for digital records, but the database was designed and installed by a person who was knowledgeable about neither the program nor the museum’s needs. It was in spreadsheet format, with fields not logically ordered for efficiency of data entry or retrieval. Search, select, sort, and copy manipulations were inefficient and frequently did not function as intended. The extended spreadsheet form was awkward and time-consuming to navigate for retrieving information.

The staff attempted to complete and clean up the ACCESS catalog, but progress was slow and the collections and related research had grown to the extent that the system did not meet data management needs. Finally, it was decided that having a data management program designed for museums would, at last, bring order and consistency to management of collections and allow enough room for complete data recording. Having the catalog consistent and up to date would permit the staff to focus on current registration and cataloging duties, facilitate access to and use of collections, and reduce staff time, effort, and wear on collections.

#### 3.1 Phase 1

A grant from the Museums for American program of the Institute of Museum and Library Services (IMLS) funded phase 1 of the digital database project, which ran from October 1, 2004 to April 30, 2006. The purpose of phase 1 was to create a new computerized database of the MOA’s collections of archaeological and ethnographic objects so that the staff and public could access useful and accurate records of objects quickly and effectively. We advertised for, interviewed, and hired a full-
time Registrar whose primary responsibility was to input data. We purchased the software and technical support necessary to create a Visual Re:discovery database based on the Collections Management module. We anticipated this would enable more efficient access to data about collections and tracking of object locations, allow for growth and expansion of collections and data, including research notes, and ultimately allow digital photographs of objects to accompany catalog information.

Visual Re:discovery is a relational database that presents data on a catalog card format so that all basic data are visible on one screen, facilitating rapid data access and retrieval. Layering of additional information screens for research and for sensitive information such as donors and object values enables them to be retrieved easily with each entry or secured to limit access when appropriate. Objects with commonalities, as well as entire collections, can be linked to an entry to facilitate expanded inquiry and research. In addition to Collections Management and Photograph Management modules, a Lexicon module assists with consistency of field terms so that indexing and retrieval will be thorough with no omitted items because of improper terminology. The lexicon can be modified to include additional categories or terms, providing flexibility yet maintaining consistency.

We mapped how to migrate existing data from ACCESS into Re:discovery, installed the new database, and migrated data onto a server at Wake Forest University. Re:discovery trained project staff in use of the new database. The Registrar input new records into the database and corrected migrated records. Project staff took tests to determine how accurately each person could perform data entry and searches and we used the results to standardize data entry procedures. We installed a public-access computer terminal with connection to the database in a museum gallery and guided members of the public and Wake Forest University students in use of the database for research and course support.

At the end of phase 1, four staff members had been trained in use of the Re:discovery database. There were 27,671 records in the database for 4,741 ethnographic and 22,930 archaeological objects. The Registrar had manually entered records for 157 ethnographic and 21,401 archaeological objects based on information on paper records. Re:discovery migrated records for 4,584 ethnographic and 1,529 archaeological objects based on objects. The Registrar had manually entered records for 157 ethnographic and 22,930 archaeological objects based on the Re:discovery database. There were 27,671 records in the database for 4,741 ethnographic and 22,930 archaeological objects. The purpose of phase 2, funded by a second IMLS Museums for America grant, was to update the computerized database of the MOA’s collections of archaeological and ethnographic objects and integrate photos of objects so that the staff and public would be able to access useful and accurate records and images of objects quickly and effectively though the World Wide Web.

Phase 1 was not a complete success, however. The public-access version of Re:discovery that we purchased with the support of IMLS was inflexible (unless we paid for customization, which we did not) and search results were visually and intellectually unexciting above the level of the individual object. This tended to limit its appeal to the public. In addition, we installed the database on a server in the Information Services building across campus from the museum. The server, intended to offer online storage for academic departments, handled large amounts of data, and frequent access, so searches were extremely slow, severely limiting appeal for researchers. Finally, we underestimated the amount of effort necessary to promote the public-access database to the public and Wake Forest community.

### 3.2 Phase 2

The purpose of phase 2, funded by a second IMLS Museums for America grant, was to update the computerized database of the MOA’s collections of archaeological and ethnographic objects and integrate photos of objects so that the staff and public would be able to access useful and accurate records and images of objects quickly and effectively though the World Wide Web.

Phase 2 began August 1, 2006 and ended July 31, 2008.

We signed a contract with Re:discovery for purchase of a Re:discovery for Internet (RFI) server software license and for custom configuration of one data directory of RFI for the museum’s computerized database. The Registrar and Instructional Technology Specialist expanded the storage capability of the existing server to hold the greatly enlarged database and worked to make the database respond more quickly to searches.

Ultimately, the Instructional Technology Specialist negotiated with Wake Forest University Information Systems to obtain and configure a virtual server with a large amount of dedicated storage capacity to hold both the private and public versions of the database and to protect them from hacking. A virtual server allowed more streamlined support because it did not entail a new, separate piece of hardware to support, and it had greater failover capabilities. The virtual server could be rebuilt on other hardware in a very short amount of time, since the configuration of the server software was not tied to a specific set of hardware. Since this new server was designed specifically to support constant access to our database, and was only used by those using our database, the performance was much improved. This new server, running Microsoft Windows Server 2003, was also a much more appropriate choice than the previous one, which was designed as a simple file server, since it offered Web server software that would support ASP (Active Server Pages) Web pages used in the RFI product. The RFI database itself was written for Microsoft SQL Server, but had also been ported to an Oracle database. Although Wake Forest University already employed a fulltime database administrator for Oracle databases, and did not have a history of supporting SQL Server, we decided to remain with the default database type used by Re:discovery because only one or two customers had installed the form of the database converted to run on an Oracle database. To cut costs somewhat, and avoid potential issues in supporting a full installation of SQL Server, we decided to run the database using the included “runtime” version of the SQL Server software. This was just the software necessary...
to run, but not edit, an SQL Server database. Since we had no interest in customizing the database ourselves, this approach resulted in much easier maintenance of the server.

The Registrar and Curator edited many existing records in the database. Student employees digitally photographed many objects; the students formatted the photos and the Registrar loaded them onto the database. By the end of the project, 102 new object records were entered, 4,360 digital photographs were made and edited, 4,300 photographs were entered into the database, and 26,419 object locations were updated from inventory records. Data cleaning and research note inputting were completed for 28,044 archaeological and ethnographic object records. The Registrar added or updated inventory information on 33,000 archaeological and ethnographic record fields based on work completed by student employees and entered records for 406 ethnographic and archaeological objects donated since the project began. These activities were in addition to those proposed to IMLS, but were possible because they could generally be completed at the same time as project activities without adversely affecting the schedule of the overall project.

During the beta-testing period late in phase 2, the public version of the database was installed on the Web. MOA staff members presented workshops for 62 local elementary and secondary school teachers to introduce them to the resource and get feedback. At last, the public version of the database (www.wfu.edu/moa/database) went live on September 9, 2008 and was announced through a reception for the public in the MOA and a press release. Since then, MOA staff members have been presenting orientation sessions for faculty in academic departments at the university and workshops for teachers through the school district office. Included within the database website are a research guide, online catalog manual, K-12 lesson plans, and answers to frequently asked questions to make it more user-friendly.

### 3.3 Phase 3

The overall goal of the phase 3, supported by a third IMLS Museums for American grant, is to provide broad public access through the Web to cataloging information and digital images for the archival collection. Access to the archival collection will enhance classroom teaching and research for faculty and students at Wake Forest University and other institutions of higher learning throughout the nation; support curriculum-based learning by students in primary and secondary schools in North Carolina and other states; facilitate research into the context of traditional material culture by scholars worldwide; and permit the general public to learn about traditions and modernization of peoples throughout the world. Project activities are scheduled to occur between August 1, 2008 and July 31, 2010.

At the onset of phase 3, it became clear that in order to maximize efficiency in updating database records, as well as to provide a consistent experience to end users, we needed to create standards for the entry and curation of digital images used in the database. These included standards for file formats and pixel sizes of digital photos and a policy for retention and storage of images. We concluded that the initial scan of paper documents, slides, and other archival materials should be of the maximum quality possible, to permit storing a copy of the material in a form as true to the original as possible. Thus, all initial scans of archival materials are stored as uncompressed TIFF (Tagged Image File Format) files. The TIFF format is “lossless,” in that it does not attempt to approximate any image data in order to produce a smaller file size. These TIFF images are stored outside the database on DVD-R discs that are stored in acid-free paper DVD sleeves, marked with an archival-quality pen on the clear hub (to avoid any damage to the reflective surface from interaction with the chemicals in the ink), and placed in a fireproof cabinet for long-term storage.

We also standardized the format and size of the images that database end users actually access. Since JPEG (Joint Photographic Experts Group) images are well-supported, are viewable by all graphical Web browsers, and provide usable images with reasonable file sizes, all images to be placed into the online catalog are converted to JPEG images, with maximum dimensions not to exceed 300 pixels in either direction. This size was tested to be usable and clear at many different monitor resolution settings, and provides an easy and attractive navigation layout for users with a variety of screen sizes. Additionally, all full-size JPEG images are saved using a 72 dpi (dots per inch) resolution, as this is the common limitation of most computer monitors. Standardization at this resolution allows us to reduce the file size of the images slightly, without limiting the quality the end user receives.

The process of capturing, converting, and storing the new digital image information for the database provides the opportunity to expand the breadth and depth of data collected and preserved. This includes preserving metadata automatically generated during the scanning or photography of the original materials, as well as inclusion of new metadata clearly identifying the images as copyrighted property of the Museum of Anthropology at Wake Forest University. All editing of images is done with Adobe Photoshop, versions CS2 and CS3, partly due to the strength of this software for batch processing of images, but also because it preserves the various metadata captured while creating the original digital image. These include the date of creation, resolution, F-stop, aperture, and other settings of the equipment used to capture the digital image. It is easy to add copyright metadata information to a batch of images during the process of converting them from TIFF to JPEG format.

All phase 3 activities are leading toward the production of a computerized database that will incorporate catalog records and digital images for each document and photograph in the museum’s archives and will be accessible to researchers, students, educators, and a broad and diverse general audience through the museum’s website. Project staff members are purchasing the Archives module in new Re:discovery Proficio software, licenses, services, and digitizing equipment. We are also upgrading from RFI to Proficio software for the Collections Management module. We are creating a digital archive of high-quality images to parallel and back up the physical archives, while simultaneously marking the physical records with catalog numbers and protecting them from damage by re-housing them for long-term storage in archival sleeves, folders, and boxes. Digital images are periodically burned to DVDs for backup. Staff members are interpreting the content of records and cataloging this information to enable access to the records. We will provide information to the software developer to configure a data directory and all software will be loaded onto the server at Wake Forest University.
Public access will be through the museum’s website. Web use manuals will be written for the museum staff, educators, and the public and four workshops will instruct 64 primary and secondary school educators on how to use the Web archives with their students. Project staff members will evaluate the project on the basis of timely completion of activities and by testing the ability of museum staff and educators to search the Web archives successfully.

4. CONCLUSION
The three phases of this project are progressively providing the staff of the Museum of Anthropology at Wake Forest University with physical and intellectual control of our object and archival collections and permitting us to fulfill our educational mission. Accurate and accessible information is contributing to formulation of a collections plan and is being incorporated into the museum’s existing disaster plan. Assessment of the condition and storage needs of the collections is being used to support requests for improved facilities and environmental safeguards.

5. ACKNOWLEDGMENTS
This project was supported in part by grants from the Institute of Museum and Library Services through the Museums for America program. Any views, findings, conclusions or recommendations expressed in this publication do not necessarily represent those of the Institute of Museum and Library Services. Wake Forest University support came from the Dean of the College, Department of Anthropology, Instructional Services, Office of Research and Sponsored Programs, and Social, Behavioral & Economic Science Research Fund. We gratefully acknowledge the contributions of MOA staff members Sara Cromwell, Tina Smith, and numerous students, and the support of the MOA Advisory Board.

6. REFERENCES
Communicating Archives of Cultural Institutions: Venice as a Case Study

Marina Niero
Ateneo Veneto Campo
San Fantin 1897 Venice 30124 Italy
niero@ateneoveneto.org

Carlo Urbani
Istituto Veneto di Scienze, Lettere ed Arti Campo
S. Stefano 2945 Venice 30124 Italy
carlo.urbani@istitutoveneto.it

ABSTRACT
In this paper, we describe the “pilot project” of archival cooperation between cultural institutions in the digitalization process.

Categories and Subject Descriptors
H.2.4. Systems, relational databases; H.3.5. Online information service, web-based services; H.3.6. Library automation, large text archives; H.4. Information system applications, workflow management; H.5.2 User interfaces, graphical user interfaces; H.5.3. Group and organization interfaces, computer-supported cooperative work

General Terms
Documentation, Experimentation, Standardization

Keywords
Ateneo Veneto, Istituto Veneto, archives, cultural institutions, cultural network, ring network

1. INTRODUCTION
After the fall of the Venetian Republic in 1797, the Napoleonic regime instituted with a decree promulgated in 1810 two different cultural institutions: one, the Istituto di Scienze, Lettere ed Arti, destined to deal with research projects and planning; the other, the Ateneo Veneto, whose function was to be a civic academy of the city, where lessons were to be held for the public by renowned professors, members of the Academy. These two institutions, although different in vocation, have shared many of the illustrious literati and scholars of Venice and the Veneto. Apart from the archival documentation accumulated in the two institutions (and in others such as the Querini Stampalia Foundation, the Saint Mark’s Library and the Correr Museum Library) throughout the two centuries to come (nineteenth and twentieth centuries), bearing witness to the fertile activity of both, many of the members decided to leave them their own papers, personal libraries or publications, creating a sort of a cultural database which explores the history of Venice and the Veneto in nineteenth-twentieth centuries from a different perspective.

2. THE PROJECT
Instead of proceeding to digitalize and catalogue the collections and archives of each institution, it was judged best to unite forces and locate an archival software platform which would answer the following criteria:

1. the final product should be a public, easy-access, online platform;
2. each institution would proceed to digitalize online its own documents, sharing common authority files;
3. each institution would have its own website, with a separate access to its own database;
4. the user would be able to navigate through the different databases, using common masks, and sharing all knowledge regarding his specific request;
5. the system will be open to other cultural institutions that could join at any moment, in order to create a sort of a regional cultural network;
6. in addition, the program should be able to offer to the user other research opportunities such as academic members’ biographies, a chronology of events occurring in the participating cultural institutions (such as lectures, debates, conferences etc.), and a link to the library collections bequeathed to the institutions, alongside the private archives, by different members.

On the whole, the two institutions decided to launch a “pilot” project using a ring network between them, like those created by CMS (Centers for Medicare and Medicaid Services) http://www.cms.hhs.gov/home/rds.asp or by London’s Museums of Health and Medicine http://www.medicalmuseums.org/museums/ana.htm, promoted by The Association of Anaesthetists of Great Britain and Ireland http://www.aagbi.org/heritage/archives.htm.

3. THE SOFTWARE
The software GEA created by the BAICR non profit consortium and developed by DATAMAT spa. is the software believed by the above cultural institutions to be the most suitable for their requirements. It is a product originally designed for twentieth-century private archives, and much used by the Region of Lazio, but also by other institutions in Italy.

The GEA software allows a flexibility of use. Its ring structure allows an interested institution to be both connected with other
servers, and stay autonomous regarding its own database. In fact every institution can decide at what level it wants to adhere to the project. It can either

1. use the software, with personal database management, undercutting costs to the minimum, as it does not acquire a “stand alone” software;
2. use the software as part of a network, with access to other databases and to the authority files.

The software also allows three other significant benefits:

1. the creation of a forum of discussion in order to solve common problems regarding cataloguing, research, etc.
2. the possibility to export data regarding the library collections to SEBINA (the Italian Union Catalogue).
3. digital curation of photographic material and the creation of links between image and text.

4. **CREATING A CULTURAL NETWORK**

“Isolated archival cataloguing” is today a worldwide problem. The decision to proceed to digitalize its own archives, may render very costly the whole operation (developing its own program or acquiring one), and the benefits are limited, as the outcome is scientifically fragmentary: it does not allow the user to grasp a fuller view of the material, scattered between different institutions, and moreover sometimes using different terminology or fields, thus creating a semantic discontinuity and the impossibility of the user to proceed to data comparability.

The intensive correspondence between different Venetian scientists and scholars with their colleagues in other cities exists, but it is scattered in different archives. Consequently, the absence of a central platform which could allow the user to “navigate” between these archives and institutions renders impossible grasping the full intensity of Veneto/Venetian cultural life in nineteenth-twentieth centuries.

One example of the importance of the “pilot project” proposed here regards the 1848 revolution in Venice. It is known that the 1848 revolution leaders, Nicolò Tommaso and Daniele Manin, members of the Ateneo Veneto, chose this institution as a center of their revolutionary activity. Yet, many other scholars adhered to their ideas, and their archives, letters, research projects, and libraries are scattered all over, between different institutions, rendering impossible at this stage to follow the cultural background that brought about the revolution.

Another example is the project “Luzzatti”, which permitted the Istituto Veneto the digitalization of all documents possessed by the renowned statesman and economist, Luigi Luzzatti (1841-1927), who was prime minister, minister of finance and one of the founders of the Italian popular banking system. Luzzatti left all his papers to the Istituto Veneto, divided in two distinct sections: the *Corrispondenza* (Correspondance) section (with 38,700 items, distributed in 4,262 fascicles, alphabetically ordered), and the *Atti* (Acts) section, constituted in 192 buste, with about 20,000 items. The project has been so far limited to cataloguing the single items without creating the archival structure, required by international standards and necessary to render it accessible online. The potentiality of the Luzzatti archives is immense: his connections were worldwide and his membership in various cultural institutions, the University of Padua and journals, if described through links between institutions, as proposed in this “pilot project” can render a new cultural dimension for Italy in 1880’-1920’.

“Communicating Archives of Cultural Institutes” intends to solve the problem of “isolated archival cataloguing” by making archival databases communicate from the start - right from the program acquisition and the digitalization step - and by creating cultural itineraries intended to provide the fullest documentation on a past scholar, member of different Venetian/Veneto institutions, who could have been in relationship with other colleagues in other Italian academies. In this manner, the user can reconstruct not only the scholar’s range of activity, but create itineraries that illustrate past virtual communities that shared common scientific and cultural interests.
The Russian Doll Effect: 
A Case Study in Digital Artifact Recontextualization

Jim Bradley
Head of Metadata and Digital Initiatives
Ball State University Libraries
Muncie, IN  47306 USA
jabradley2@bsu.edu

ABSTRACT
This paper explores a specific project to digitize and make available artworks of the Ball State University Museum of Art. By establishing partnerships and maintaining flexible metadata, the portability and recontextualization of digital artifacts (termed by the author as the ‘Russian Doll Effect’) has been maximized. The case study details the primary context of the digital assets, current recontextualizations, and future directions.

Keywords
Digitization, partnership, art, digital curation, Second Life, metadata

1. INTRODUCTION
A well-planned digital artifact will find itself repurposed and repackaged several times for a variety of environments, gateways, and diverse user types – a phenomenon I refer to in recent writings [1] as the ‘Russian Doll Effect’ (in reference to matryoska, a popular toy consisting of several dolls nested within one another) [2].

In this paper, I present a case study that illustrates this effect, detailing an ongoing digitization project that has evolved beyond its original intent – discussing both present and future plans for artifact recontextualizations.

2. A SIMPLE COLLECTION
In December of 2004, I joined Ball State University Libraries specifically to create the Digital Media Repository (http://libx.bsu.edu). Now, less than four years later, this digital collection has grown to contain more than 100,000 individual items. The success of the Digital Media Repository (DMR) has been largely due to partnerships with granting agencies, museums, historical institutions, and many of the university’s various academic departments.

One of these partnerships was with the Ball State University Museum of Art – who approached us soon after receiving a grant from the Institute of Museum and Library Science (IMLS) and the Ball State University Enhanced Provost Initiative to create high-quality digital images of the 11,000 artworks in the Museum’s collection. The Museum had no need for assistance in photographing the artwork; rather, the end goal of the project was to place the images in an online searchable database, and the Digital Media Repository seemed a good fit.

While the photographer began working his way though the artworks, a long series of meetings between the Museum and the Library were held, slowly and carefully establishing the range and syntax of a metadata schema that would fulfill the needs of the Museum, the Library, and (most importantly) the end user.

Other, less technical issues such as branding, page design, and collection gateway flexibility were negotiated – eventually satisfying both parties – and with this compromise, the initial collection of 100 images went live in the spring of 20081 – with the goal of expanding the collection to approximately 1,000 images by 2009.

3. DISCOVERING A ‘SECOND LIFE’
In originally unrelated projects, John Fillwalk – director of the Intitute for Digital Intermedia Arts and Animation (IDIAA) and associate professor of art – had been working to construct a presence for Ball State University in the Second Life interactive virtual environment.

A common critique of the Second Life environment is that, in a practical sense, it has broken very little new ground. Although the use of an avatar and a three-dimensional environment is quite novel, it is essentially a more cumbersome (albeit perhaps more entertaining) method of performing tasks that could be more readily achieved through a web browser. This perception is reinforced by the fact that, when Second Life provides access to an online resource, it is usually served by opening a traditional web browser window.

The plans for a Second Life version of the Ball State University Museum of Art progressed. It was to feature a ‘virtual art gallery,’ - a three dimensional environment in which one’s avatar could peruse the artworks of the Ball State University Museum of Art.

This work is licensed under the Creative Commons Attribution-Noncommercial-No Derivative Works 3.0 Unported license. You are free to share this work (copy, distribute and transmit) under the following conditions: attribution, non-commercial, and no derivative works. To view a copy of this license, visit http://creativecommons.org/licenses/by-nc-nd/3.0/.

DigCCurr2009, April 1-3, 2009, Chapel Hill, NC, USA

1 The collection is accessible through the Ball State University Digital Media Repository at (http://libx.bsu.edu/collection.php?CISOROOT=/MuseumBSU) or through the Ball State University Museum of Art DIDO Project gateway at (http://www.bsu.edu/artmuseum/dido/)
However, when the discussion turned to accessing the Museum of Art Materials in the Digital Media Repository, a possible new paradigm emerged: rather than using the virtual art gallery as a stepping-off point to the collection, perhaps we could allow the user to directly query the DMR metadata from within the Second Life environment – thus populating the ‘walls’ of the ‘gallery’ with artworks in accordance the user’s search results.

4. FUTURE DIRECTIONS

If we redefine digital librarianship as digital stewardship, we take upon ourselves the mission to create and manage digital objects that are adaptable and sustainable over the longest period possible. Accordingly, we must assume that today’s digital objects will be subject to the ‘Russian Doll Effect’ – that they will be utilized outside of their original context, repurposed and embedded within secondary environments, and accessed by a diverse user group using a variety of ever-changing information pathways and technologies.

Thus, as we move toward the launch of the Second Life museum, I try to imagine more new and interesting ways to reuse and repurpose our digital assets. In closing, I would like to briefly outline two future directions for the Ball State University Museum of Art Image Collection that are under consideration.

4.1 The ‘Pocket Tour Guide’

A simplified derivative version of the DMR/DIDO collection could be distilled from the XML via XSLT, designed to be accessible via mobile devices and cell phones. Thus, Museum of Art visitors could instantly access additional information on any piece of art – discovering the provenance of the piece, the materials or techniques used in its construction, or a brief biography of the artist.

4.2 The ‘Lesson Plan’

Working in conjunction with the Ball State University Museum of Art Curator of Education, images from the collection could be identified and embedded within instructional packets. Educators could easily locate and download age-appropriate materials that conform to curriculum guidelines. The ability to offer the images and lesson plans via the internet would be of great interest – particularly at schools in which art and music programs are endangered due to funding restrictions.

5. ACKNOWLEDGMENTS

I would like to thank Ball State University Libraries, the Ball State University Museum of Art, the Ball State University Center for Media Design and the Institute for Digital Intermedia Arts and Animation for their collaboration on the projects discussed in this paper. Additional thanks to the Institute of Museum and Library Sciences (IMLS) and the Ball State University Enhanced Provost Initiative for their generous support of these projects.

6. REFERENCES


[2] It should be noted that when I began using the Russian Doll metaphor in 2005, I was unaware that fellow scholar Günther Waibel used the same analogy in his 2003 article: “Like Russian Dolls: Nesting Standards for Digital Preservation” (RLG DigiNews 7:3) – thus, all due credit for the original coining of the term should be extended to him.
ABSTRACT
This paper presents a framework that was developed from a study in 2008 of archival education in Mexico. Based on the results of survey data, semi-structured interviews, and ethnographic data, a framework was developed consisting of six elements (conceptual expansion; embeddedness; collaboration; leadership, activism and ethics; sustainability; and reflexivity), which are useful for capturing a plurality of perspectives when developing a culturally sensitive graduate-level curricular framework and course modules to prepare students for digital curation in various environments. A case study of the applicability of the framework within digital curation education is presented to illustrate further the necessity of this conceptual approach.

Keywords
Digital archives, pedagogy, cultural sensitivity, digital curation education, digital curation,

1. INTRODUCTION
Digital technologies have reached a level of pervasiveness that has begun to fundamentally alter how society interacts with information. Technological infrastructures play a mediating role in how people, objects, and information interact. Increasingly in museums, libraries, and archives, digital technologies play significant roles in preserving cultural heritage knowledge and materials. Those technologies not only allow archives, libraries, and other information institutions to expand by developing digital repositories, but also to distribute information over networks; employing broadened notions of “freedom of information” and “access to knowledge.”

Although “advancement” in technological infrastructures has occurred, little progress has been made in developing digital information systems in ways that incorporate the interests, needs, and diverse cultural beliefs in Indigenous, ethnic, and other minority communities that were precluded from full participation in society due to the legacies of slavery, colonialism, imperialism, genocide, as well as phenomena such as racism, ethnocentrism, and homophobia. Digital technologies often serve as platforms to produce and capture cultural heritage knowledge and materials; however, if developed and employed in ways that are culturally irrelevant or insensitive, those technologies, though “advanced,” may function as mere extensions of socio-cultural barriers, which often exist between mainstream and subaltern communities. Culturally relevant digital technologies, whether utilized by archives, museums, or libraries, have the potential to function as counternarratives by providing opportunities to tell more stories from different perspectives, and to support new types of users. In some situations, culturally sensitive technologies may actively contribute to the empowerment, examination, and redress of subaltern communities.

2. FRAMEWORK
This paper presents a framework that was developed from a study in 2008 of archival education in Mexico [1]. The study’s overall goal was to understand Mexico’s archival education infrastructure. In doing so, it sought to provide insight on how communities of African heritage became absent from Mexico’s official record; to understand the role that education of archival professionals might play in addressing or contributing to these absences; and to generate recommendations for how underdocumentation might be partly remediated by changing what is currently taught in formal archival education at the university level. Based on the results of survey data of archival educators and practitioners, semi-structured interviews of cultural gate-keepers in the Afro-Mexican community, and ethnographic data of the ways of remembering in Afro-Mexican communities, a framework was developed consisting of the following six elements:

1. Conceptual expansion, which addresses different conceptualizations of the record by different communities, particularly those with non-Western epistemologies;

2. Embeddedness, which addresses locating field experiences within communities to gain a richer understanding of community needs;

3. Collaboration, which addresses partnering with community-based organizations in efforts to cultivate equitable, mutually beneficial partnerships;
4. **Leadership, activism and ethics**, which addresses expanding the archival role in promoting visibility of under-documented communities;

5. **Sustainability**, which addresses planning that is sensitive to the community’s resources and relevant to its cultural protocols; and

6. **Reflexivity**, which addresses critical examination of the body of knowledge comprising archival theory and practice, but also the role the academic

Those elements are useful in systematically incorporating the interests, needs, and cultural beliefs of diverse communities into Mexico’s archival education curriculum. This framework is also useful in developing a culturally sensitive graduate-level curricular framework and course modules to prepare students for digital curation in various environments.

### 3. CASE STUDY

To demonstrate the applicability of the above framework within digital archival development and digital curation education, we present the following case study. First the context of the study will be presented, followed by examples on how the framework applied to each scenario.

#### 3.1 Case Study Context

This case study describes the development of a digital archive for the Buffalo Fire Historical Society Museum (BFHSM) during Spring 2006. The class included 30 students in the Library and Information Studies master’s program at the State University of New York at Buffalo. The plurality of perspectives included in this case does not pertain to ethnicity or gender per se, as the study described above does. This case addresses differences in intellectual, professional, and community culture found among users of the physical museum (including the volunteers who staff the museum, themselves active or retired firefighters, local firefighters and their friends and families, and other firefighters and fire buffs who access the collection online). The objects of the collection included items from the BFHSM that were chosen by the volunteers who staff the museum as representative samples of their “remembering” or to display the history of their culture. Some of these items included: fire call boxes, fire hydrants, leather fire buckets, photographs and historic postcards, scale models of fire boats and fire engines, and firefighting equipment (antique fire engines, fire carts, hose reels, etc.) “Remembering”, or representing their culture’s history, for the firefighters was evident in the objects they chose to include in the digital archive, as well as the objects they displayed publicly in the museum. Firefighters not only have specialized artifacts that are used in the course of their activities, but they also have a unique language they use to talk about and to create records of the objects in their collections.

#### 3.2 Case Study Context

This case study describes the development of a digital archive for the Buffalo Fire Historical Society Museum (BFHSM) during Spring 2006. The class included 30 students in the Library and Information Studies master’s program at the State University of New York at Buffalo. The plurality of perspectives included in this case does not pertain to ethnicity or gender per se, as the study described above does. This case addresses differences in intellectual, professional, and community culture found among users of the physical museum (including the volunteers who staff the museum, themselves active or retired firefighters, local firefighters and their friends and families, and other firefighters and fire buffs who access the collection online). The objects of the collection included items from the BFHSM that were chosen by the volunteers who staff the museum as representative samples of their “remembering” or to display the history of their culture. Some of these items included: fire call boxes, fire hydrants, leather fire buckets, photographs and historic postcards, scale models of fire boats and fire engines, and firefighting equipment (antique fire engines, fire carts, hose reels, etc.) “Remembering”, or representing their culture’s history, for the firefighters was evident in the objects they chose to include in the digital archive, as well as the objects they displayed publicly in the museum. Firefighters not only have specialized artifacts that are used in the course of their activities, but they also have a unique language they use to talk about and to create records of the objects in their collections.

#### 3.3 Pedagogical Activities of Project

Students worked in teams to develop both the technical and intellectual infrastructures of the digital archive. There were seven directed teams that were embedded into the culture of the BFHSM. Students spent 3.5 class hours per a week in the museum, but also devoted time outside of class to working at the museum in order to complete their team activities. Each team worked closely with the volunteers in order to learn about the culture’s:

- users (all populations present both in house and online),
- uses of the collection (information needs or requests submitted by users via phone or email),
- significance of the objects in the museum and why they were chosen for display,
- scope, historical significance, and provenance of the collections,
- copyright or use policies associated with the collection’s objects but also with donating objects or using objects for digital display,
- technical knowledge and computing and digitization skills of the volunteer staff and other local member users,
- technical equipment present in the museum.

#### 3.4 Resulting Findings and Products

The result of the students’ work with the volunteer staff informed the development of the technical infrastructure (choice of infrastructure needed, database design, search functions of digital archive, and interface to digital archive) and intellectual infrastructure (collection development and copyright and use policies; digitization standards for digital photography, digital audio, and digital scanning captures as well as quality assurance and work flow documents; metadata schema customization; choice of indexing language(s); and input standards for creating
leadership, activism, and ethics: As spent many hours at the museum with the volunteer staff. To learn more about the culture they attended member meetings of the Buffalo Fire Historical Society, as well organization. To learn more about the culture they attended each object had a story to tell. Also, as most of the items in the collection were three dimensional objects, differing spatial views were needed to accurately depict the objects as records of their remembering.

Collaboration: Students partnered with a community-based organization. To learn more about the culture they attended member meetings of the Buffalo Fire Historical Society, as well as spent many hours at the museum with the volunteer staff.

Leadership, activism, and ethics: In order to design a “culturally sensitive” digital archive, the students needed to gain insight into the cultural identity of the BFHSM and its users. It was extremely important to them to accurately represent the culture in both the technical and intellectual infrastructures of the digital archive. Their work contributed to the “identity” and promoted visibility of the firefighter culture.

Sustainability: The issue of sustainability of a digital archive is one that is both a challenge and a learning opportunity for digital curation students. Prior to beginning a digital archive project questions about resources and plans for sustainability must be asked. Because many museums are nonprofit organizations and have limited budgets, choices regarding technology used and skills needed by the members of the culture must also be culturally sensitive. In this project the choice of infrastructure was determined based on conversations with the museum staff. They had both a limited budget and low level technology skills. One member of the staff had more advanced technology skills and was the webmaster for BFHSM. She also had experience with database development, and various digital capture techniques, including photography, and scanning. Cognizant of these limitations, the student teams chose to use technologies that would allow the BFHSM to sustain the digital archive with minimal cost and low technical expertise. It was also necessary to develop very detailed work flow documents and digital capture manuals and quality assurance instruments, as well as comprehensive, but easy to use collection development and copyright/use policies.

Reflexivity: This project-based experience allowed both the students and the instructor the opportunity to re-examine their roles as educators and contributors within the culture of firefighters, as well as within the larger community of Buffalo, NY. It was important for the instructor not to assume any hierarchical role in the project, other than project manager, but also by necessity, to act in the role of facilitator, trainer, and mentor to students and the staff of the BFHSM. Because of the perceived power structure between the “educated” and community members, students and the instructor (by mere fact that they had more intellectual tools to work with, and thus, were probably perceived as having more power) were constantly aware that it was detrimental to assume that they knew and understood the object of knowledge only from their own cultural or intellectual viewpoint. Taking a dialogic approach, which requires reflexivity, students were able to gain cultural competency of the firefighter community.

4. CONCLUSION
The authors believe that by implementing the above framework, the interests, needs, and diverse cultural beliefs of non-mainstream communities will be addressed in culturally appropriate ways as digital information systems and graduate level curriculum are developed. Doing so will create opportunities for educators and students to gain a richer understanding of diverse user communities and their interaction with all forms of records, archives, museums in and across space and time, while addressing the challenges of differing cultural understandings and limited technological knowledge and access.

5. REFERENCES
ABSTRACT
A debate at DigCCurr 2009 will allow educators and practitioners the chance to reflect on the balance of practical and theoretical skills required of digital curators. The area of significant properties is used as a case study for drawing out some of these skills in research and teaching environments. A panel will debate issues, with the audience invited to contribute to the discussion.

General Terms
Management, Theory

Keywords
Significant properties, pedagogy

1. INTRODUCTION
Roles and responsibilities of information professionals are changing in the distributed digital environment [1]. The Centre for e-Research (CeRch) at King’s College London is planning an innovative postgraduate course in Digital Asset Management for 2009. This will seek to balance the practical skills and techniques of digital curation with a more theoretical approach, to offer students a rich mix of course content able to sustain the future working lives of reflective practitioners. Similarly, the DigCCurr and DigCCurr II projects at the School of Information and Library Science at University of North Carolina at Chapel Hill, are building an international digital curation curricula and professional engagement opportunities for graduate students and practitioners.

A research project underway at CeRch exemplifies digital curation’s mix of skills. InSPECT is investigating the significant properties of digital objects over time [2]. Significant properties can be seen as a useful arena to discuss issues of theory and practice across information, library and archival sciences.

2. SIGNIFICANT PROPERTIES
Significant properties (SPs) are those characteristics of digital objects that should be preserved over time in order to perpetuate their “quality, usability, rendering, and behaviour” [3]. SPs relate to the information object in the OAIS data model, in capturing the content, context, appearance/rendering, structure and behavior of objects. SPs are related to and sometimes overlap with what the Reference Model for an Open Archival Information System (OAIS) [4] calls Representation Information (RI). RI relates to the information needed to make sense of the data object as SPs do at the level of the information object.

Although tools may be able to discover preserved digital objects, making meaningful use and sense of such digital objects will also require contextual information [5]. SPs, when identified and preserved in metadata with digital objects, offer evidence of the wider context (beyond the datastream itself) needed by future users. In addition to the SPs of individual digital objects, other contextual information -- such as on provenance, authenticity or relationships to other digital objects -- may also be crucial for users to make sense of digital objects. Four skills and their academic parentage are outlined as important for those engaged with SP research and practice. These will form the basis of a discussion.

3. SIGNIFICANT SKILLS
3.1 Characterization
Characterization supports the logical accessibility of digital objects over time by identifying the precise format version of an object and validating that object against relevant formal specifications. The activity comes from computer science, and requires an understanding of how digital files are structured and encoded. Analyzing files to enumerate SPs can be a lengthy, involved process, although tools such as JHOVE can support this activity to some extent.

3.2 Diplomatics
The archival practice of diplomatics is concerned with the study of the genesis and form of individual documents, to determine if they are what they purport to be. This culture of analysis and judgment determines the authenticity and integrity of records, and has been applied to the digital domain. Digital curators will need to be at least as confident of the authenticity and integrity of their content as traditional archivists, since the technical means of establishing direct links with the creators and their methods may be lost or partial.
3.3 Epistemology
Epistemology is the branch of philosophy concerned with knowledge. Social epistemology, in particular, can help to describe the social dimensions of information. The content of digital objects conveys knowledge as intended by the creator. Librarians and archivists have not always had subject skills to verify or warrant the reliability of information conveyed in materials they manage, as opposed to their authenticity. In the digital domain, the respective contexts of the creator and that of the user can be far apart unbound by ties of culture, intellectual parity, time or space.

3.4 Conceptualization
Digital humanities (or humanities computing) as a discipline provides a methodological focus for studying the outputs of human creativity. It also has an understanding that increasingly computers are used as ‘venues of representation’ rather than just as computational machines. As people become more adept at using technology to create digital forms of representing information and meaning, curators require a similar familiarity with the possibilities of human interaction with such forms. Content modeling, visualization and other forms of conceptualization inform the description and understanding of digital content, and can offer entirely new approaches to the reuse of data.

4. PRACTICAL VS. THEORETICAL
An academic discipline needs to define the limits of its domain, the identifiable body of knowledge which an adherent can comprehend. It also needs to be able to conceptualize that domain. Librarianship and archival science have long articulated their respective domains’ knowledge bases, and also their traditions of practice. The role of teaching practical skills plays can be problematic in the academy, since practice changes over time in response to new research, changing technology and human factors. Yet communities of practice also need to have embedded a shared theoretical understanding for the vision to adapt to the changing environment.

The ‘performance model’ articulate by the National Archives of Australia [6] is a useful way of looking at significant properties. It combines the source datastream and process of rendering in a performance that understands the human interaction with technology. It offers a useful grounding for the digital curator’s task in giving life to digital objects over time.

5. DEBATE
A multinational panel will be convened to debate the mix of skills required of digital curators, the roots of such skills and how they might develop in coming years. A short introductory statement will frame the debate by using significant properties as a case study combining practical and theoretical skills. The audience will have a sense of digital curation as an evolving domain and field of activity, with a vital role in the enduring availability of scientific and cultural value, and not just the half-born offspring of other external disciplines.

6. REFERENCES
Invited Panel: Change Management

Fynnette Eaton
POST
Electronic Records Archives Program (ERA)
College Park, MD 20740 USA
jemfle@aol.com

Jonathan Crabtree
Odum Institute for Research in Social Science
Univ. of North Carolina at Chapel Hill
Chapel Hill, NC 27599 USA
jonathan_crabtree@unc.edu

Stephen Chapman
Harvard University Open Collections Program
Cambridge, MA 02138 USA
stephen_chapman@harvard.edu

ABSTRACT
This invited panel on change management in the digital curation environment. It will have three speakers representing different perspectives. Fynnette Eaton served as the Change Management Officer for the Electronic Records Archives Program at the National Archives from 2002-2007 and will provide an overview of types of issues that anyone introducing major changes in an organization will face. She will use examples from her experiences both at the Smithsonian and the National Archives. The emphasis will be on dealing with a system that will change how staff performs its work. Stephen Chapman from Harvard University will discuss the experiences in the Open Collections Program at his institution as they deal with changes in digitization mandates that will change how staff performs its work. One example is the shift from a mass digitization mandate (for published materials) to one that focuses upon production digitization for unique materials (archives, manuscripts and rare books) which required across-the-board adoption of workflows and systems. All of which underscores the point that adaptability and other skills are intrinsic to the ongoing success of digitization programs. Jonathan Crabtree will share some experiences that Odum has had during the DataPASS project and migration to the current Dataverse archive distribution and management software. The major change being the shift to a federated approach for preservation and the IT infrastructure that comes with it. Changes in workflow are important but in addition it is the social structures of the federated environment and the efforts placed on partnership building. This has had significant impact on collection development and the acquisitions process. In addition the shift in user patterns causes concern and consequences in the collection of usage statistics. These changes require attention on both the staff development front as well as administrative expectations.

Keywords
Digital Curation; Digital Preservation; Change Management.

This work is licensed under the Creative Commons Attribution-Noncommercial-No Derivative Works 3.0 Unported license. You are free to share this work (copy, distribute and transmit) under the following conditions: attribution, non-commercial, and no derivative works. To view a copy of this license, visit http://creativecommons.org/licenses/by-nc-nd/3.0/.

DigCCurr2009, April 1-3, 2009, Chapel Hill, NC, USA
An Implementation of the Audit Control Environment (ACE) to Support the Long Term Integrity of Digital Archives

Michael Smorul
Institute for Advanced Computer Studies
University of Maryland, College Park
College Park, MD 20742 USA
toaster@umiacs.umd.edu

Sangchul Song
Department of Electrical and Computer Engineering
Institute for Advanced Computer Studies
University of Maryland, College Park
College Park, MD 20742 USA
scsong@umiacs.umd.edu

Joseph JaJa
Department of Electrical and Computer Engineering
Institute for Advanced Computer Studies
University of Maryland, College Park
College Park, MD 20742 USA
joseph@umiacs.umd.edu

ABSTRACT
In this paper, we describe the implementation of the Audit Control Environment (ACE)[1] system that provides a scalable, auditable platform for ensuring the integrity of digital archival holdings. The core of ACE is a small integrity token issued for each monitored item, which is part of a larger, externally auditable cryptographic system. Two components that describe this system, an Audit Manager and Integrity Management Service, have been developed and released. The Audit Manager component is designed to be installed locally at the archive, while the Integrity Management Service is a centralized, publically available service. ACE allows for the monitoring of collections on a variety of disk and grid based storage systems. Each collection in ACE is subject to monitoring based on a customizable policy. The released ACE Version 1.0 has been tested extensively on a wide variety of collections in both centralized and distributed environments.

Categories and Subject Descriptors
H.3.4 [Information Storage and Retrieval]: Systems and Software; H.3.7 [Information Storage and Retrieval]: Digital Libraries

Keywords
ACE, Data Integrity, Digital Archiving.

1. INTRODUCTION
In this paper, we introduce a general software environment called ACE (Auditing Control Environment), which is based on a rigorous cryptographic approach and yet quite efficient and can interoperate with any archiving architecture. Using the new framework, we introduce procedures to continually verify the integrity of the archive. Our approach will allow an independent auditor to verify the integrity of every version of an archived digital object as well as link the current version to the original form of the object when it was ingested into the archive.

Specifically, ACE is based on creating a small-size integrity token for each digital object upon its deposit into the archive (or upon registration of the object of an existing archive), to be stored either with the object itself or in a registry at the archive as authenticity metadata. Cryptographic summary information that depends on all the objects registered during a dynamic time period is stored and managed separately. The summary information is very compact and is of size independent of the number or sizes of the objects ingested. Regular audits will be continuously conducted, which will make use of the integrity tokens and the summary integrity information to ensure the integrity of both the objects and the integrity information. In our implementation, audits can also be triggered by an archive manager or by a user upon data access. However we are assuming that the auditing services are not allowed to change the content of the archive even if errors are detected. The responsibility for correcting errors is left to the archive administrator after being alerted by the auditing service.

2. Overview of ACE Integrity Approach
ACE adopts a two-tier approach. The first tier deals with creating a small size Integrity Token (IT) (Figure 1) for each digital object upon its deposit into the archive (or upon registration of the object of an existing archive), to be stored either with the object itself or in a registry at the archive as authenticity metadata. Cryptographic Summary Information (CSI) depending on all the objects registered during a dynamically adjustable time interval is stored and managed independently of and separately of the archive. The ITs and CSIs are used to continually verify the authenticity of the corresponding digital object. The second tier involves the generation of very compact witness values that cryptographically depend on all the objects ingested during the previous day.

2.1 Tier 1
The first tier integrity information types (IT and CSI) are generated in two steps; aggregative registration and hash-linking. The aggregative registration of the objects is typically invoked during ingestion, and composed of aggregation rounds. The
interval of each round is determined dynamically based on the number of registration requests and time passed. This dynamic aggregation period allows us to control both the maximum size of ITs and maximum wait-time for registration. During an aggregation round, the hashes of all the objects submitted for registration as well as random hashes as necessary are aggregated using an authentication tree such as the Merkle’s tree [2]. Note that, in practice, the hash of the object is submitted as a part of the registration request (IT Req in Figure 1). The internal node in the authentication tree has the hash value of the concatenated hashes of the children.

We insert random hash values into each round to ensure that the tree will always have a certain minimal number of leaves. Figure 2 shows an authentication tree for a round involving eight objects with hash values $h_0, h_1, h_2, ..., h_7$.

2.2 Tier 2

As mentioned before, the second tier deals with generating witness values that will ensure the integrity of CSIs which are generated from the first tier operations. A witness value is constructed by aggregating the CSIs that have been created over each day, using an authentication tree whose root value becomes the witness value of the day. These witness values are published over the Internet at well-known public sites offering storage, library, or publication services. Since these witness values are small in size (less than 100KB a year), we also store them on a CD ROM (in fact, on multiple CD-ROMs that are refreshed on a regular basis). Printed versions of this witness are also possible as one line per witness would only require around 30 pages of paper for an entire year! ACE currently uses the Internet newsgroups at Google and UMIACS to publish witness values.

3. ACE Workflow

Two different workflows have been implemented in the first release of ACE. The first is a token registration workflow where new Integrity Tokens are issued from an IMS. The second workflow is the validation workflow where previously issued tokens are used to validate the integrity of files and the token itself.

Registration and validation is performed by an Audit Manager (AM). This audit manager runs physically close to the data that is to be monitored. It is designed to have bit-level access to the data so that it may read all monitored files and generate digests across those files.

The AM requests ITs from an Integrity Management Service (IMS). The IMS performs round aggregation and witness generation as described above. In addition, it also acts as a remote repository for witness values and CSIs. Of course, to fully audit the IMS, a 3rd party would use their own copy of the witness value.
3.1 Token Issuing

Tokens are issued as part of the first tier described earlier. The AM generates a SHA-256 digest of the file to be monitored. This generated digest and file name is submitted to the IMS for inclusion in the current round.

The submitted token is aggregated with other requests during the same time interval. The resulting CSI is stored in a database to be later used for witness generation and IT validation. For each request, an IT is generated and returned to the client. This flow is shown in Figure 4.

3.2 File and Token Validation

File and token validation occur on the AM subject to a specified policy. This policy may vary between collections. The AM locally stores a copy of digests for each item in a collection. Periodically, each monitored item is read, a digest is calculated and compared with the stored value.

Token validation requires showing the stored digest has not been altered. Validation is done by linking the stored digest to an IMS CSI. This involves calculating the round CSI using the authentication tree stored in the IT and the stored item’s digest. The IMS is then queried to retrieve the CSI for the round where the IT was issued. The returned CSI is compared to the calculated CSI and if it matches, the IT and digest are considered trustworthy to a high probability.

4. System Components

Version 1.0 of ACE has seen the release of three components. The first is an implementation of the Integrity Management Service which performs round aggregation, token issuing and witness publication. Second is a Java programming interface that allows for high performance access to IMS functionality while being simple to use. Third is a web-based Audit Manager designed to be installed by individual archives to monitor their collections.

Of these components, most archives will only need to concern themselves with managing a local Audit Manager. The IT requirements for installing an AM were designed to be minimal, requiring only MySQL, Java, and Apache Tomcat.

4.1 Integrity Management Service (IMS)

The IMS issues client tokens, stores CSI round data, and publishes a witness value each night. The IMS has been implemented as a Java EE application that provides all IMS functionality as a set of web services. A publicly available IMS is running at the University of Maryland at the address ims.umiacs.umd.edu.

The IMS supports two types of aggregation rounds. First is a timed round that will terminate once a timeout or request threshold is reached. The second round is an immediately generated round requested by a client. The immediately requested round will cause the IMS to close any open round, perform aggregation and issue Integrity Tokens.

The IMS offers several web services that allow tokens to be issued and validated. These calls are described below:

- **requestTokensAsync** – A non-blocking request for a token or multiple tokens. A receipt is returned to the client that can be used to later retrieve the issued tokens. Requests are added to the current round.
- **retrieveTokens** – Called after requestTokensAsync to retrieve the tokens responses. This must be called after the current round closes or an error will be returned to the client indicating tokens have not yet been calculated.
- **requestTokensImmediate** – A blocking request for a token or multiple tokens. The call requests the tokens be added to the current round and the round be closed. The call will return the requested tokens. The round may include hashes other than ones requested in the call if previous calls to requestTokensAsync were made in the current round.
- **getRoundSummaries** – This call returns a list of CSIs for the requested rounds. This will be used by clients to verify the integrity of its tokens and hashes.

The IMS is backed by a MySQL database that stores unclaimed token responses, round summaries, and nightly witness values.

Nightly witness publication is handled through an API that allows for pluggable publication methods. The WitnessPublisher API provides an abstract class that additional publication methods...
must extend. The IMS currently supports an e-mail based publication service.

The IMS installed at umiacs has two e-mail targets configured. First is a listserv hosted at UMIACS called ims-witness available at: http://mailman.umiacs.umd.edu/mailman/listinfo/ims-witness. Second, a Google group has been created called ace-ims-witness which is available at http://groups.google.com/group/ace-ims-witness. Both mail lists are publically available, allowing anyone to subscribe and receive nightly witness reports. In addition, both lists archive all published witness values.

Archives that wish to challenge the integrity of the IMS in the future should subscribe to one or both of these lists. While public archives of past witness values are available, the values may be considered more trustworthy as an archive can show the provenance of a witness value from publication onward.

4.2 IMS API

A Java API has been written to allow for easy high performance communication with the ACE IMS. The API provides a multithreaded library that allows a client to serially request tokens or token validation while batching those requests and transmitting them in a separate thread. Results are returned to a client through a registered callback. This allows clients to use the bulk request ability of the IMS without having to rework their process to account for the batch processing.

The core of the IMS API is the IMSService class. This class provides a connection to all IMS functionality as well as factory methods for creating token request and validation processes. The token request and validation processes consist of two parts. First is a queue that clients can serially add requests into. The queue will accumulate requests until either a maximum queue size or timeout is reached. Once either condition occurs, a background thread will be notified. This background thread will copy the work queue and send a request to the IMS. During this process, the client is free to add items to the queue. When the IMS response is complete, the background thread will notify a client supplied callback of the IMS response. Caution must be taken by the client to ensure the callback it supplies is thread-safe with respect to the client thread.

4.3 Audit Manager

The ACE Audit Manager is a Tomcat based Java web application that actively monitors collections on a variety of storage resources. The AM provides a simple web-based dashboard view of all collections that are stored in an archive. After installation of an AM, management of collections is designed to be handled by archivists rather than local IT administrators. A single AM is able to monitor multiple collections on a variety of storage platforms.

An AM handles both registration of new items and monitoring of existing items. The AM is able to request tokens for new items in collections, validate items against their stored digests, and verify those digests using integrity tokens and the IMS. Each collection is able to specify a different audit policy. It also provides complete logging of all actions performed against a collection as well as extensive browsing and reporting capability.

4.3.1 Design

The Audit manager is designed to support multiple types of storage. To do this, it must make a few assumptions about the types of storage it will be operating on. First, this storage is hierarchical. This is generally not a problem, as most filesystems and storage systems are hierarchical in nature. Second, all items are discrete objects. Objects should not be compound objects.

The audit manager stores items in a collection, organized by the parent/child relationship of the items. This allows administrators to browse collections in the same way they would browse files on a hard drive. For each item, the following information is stored

The Audit Manager has a Service Provider Interface (SPI) that allows drivers for other storage mediums to be added. While the AM is able to store all integrity information, it requires the driver to provide a gateway to the underlying storage system and handle connection specific information.

Using the SPI, we have implemented interfaces to the Storage Resource Broker[4], the Integrated Rule-Oriented Data System[5], storage local to the Audit Manager server, and a benchmarking driver to determine maximum performance of a particular AM installation.

4.3.2 Collection Registration and Audit

Collection registration involves gathering all necessary information needed for the AM to communicate with the underlying storage. After a collection has been registered, the AM will perform the first audit of a collection in which it generates tokens for all items in that collection. The following steps illustrate the registration process.

1. Display registration page to client requesting collection root and storage type/driver.
2. Display additional configuration parameters (username, password) as required by the driver. Validate all configuration parameters against the driver.
3. Start a collection audit. Request the driver supply a list of all items in a collection
4. For each item in the collection, register the new item in the database, marking it active, but missing a token.
5. Use the IMS Api to request a token for the new item’s hash.
6. Receive notification that a token has been issued.
7. Register new token in a database.
8. Finish Audit.

4.3.3 Collection Metadata

During the auditing process a large amount of metadata is collected for each collection. Each item in the collection has several metadata elements stored to perform audits. In addition, each audit, additional metadata is generated in the form of event logging. These events track every state change of an item while it is under the supervision of ACE.

For each item, the following metadata attributes are collected:
Item path – Complete path relative to the root of the collection
First Seen – Date this item was added to the audit manager.
Last Seen – Last time this item was read and validated. For the initial registration, this will be the first seen date.
State – Current state of the file. Can be one of the following:
- A – Active, item is intact, readable and digests match
- C – Item is present, but its digest doesn’t match the stored value
- M – Item is missing or cannot be fully read.
- T – Item is registered, but a token has not been received yet.
Change Date – Date the item’s state was last changed. For new items, this will be the date a token was issued.
Token – Token containing digest and IMS response.

A number of events that change the state of items in a collection will be encountered. These different events have been classified and are recorded in the audit manager. Each event contains the following information:
Event Type – The type of event that occurred. Currently, there are 19 different types of generated events.
Description – detailed description containing any error message or other information that caused this event
Session – Audit session this event occurred in
Date – Date this event occurred.

Events in the AM are grouped by a session identifier. This session identifier connects an event with other events that occurred during the same audit. If only one audit pass has been performed on a collection, then all events associated with that collection will have the same session id. Sessions allow viewing of any events that occurred during a given audit.

Audit messages can be grouped into two broad categories, normal operating messages and error messages. Normal messages show new files and tokens being added to a collection while error messages show corrupt files, tokens, or storage errors.

Managers are able to filter by item path, collection, event category, and session. This allows for complex queries such as ‘show all events for file X in this session’ or ‘show all errors ever registered for a collection’.

4.3.4 Browsing and Collection Reporting
The Audit Manager provides for collection browsing and reporting. The ACE browser provides a file-system view of items stored in a collection. Items are browsed by expanding folders and clicking on files to view details. From the Browse interface, it’s possible to view a token issued to a file, remove a file or directory, download the content of a file, and view the event log for a file.

The Audit Manager is able to generate a report showing items that are corrupt, missing, or otherwise not intact. From this report, managers are able to view log entries associated with flagged items or remove the item from monitoring if the item is later deemed correct. Removal and re-registration of items may be necessary if a new version of an item was added to the archive.

Reports comparing collections can also be generated. For example, a depositor may have a list of digests and filenames they believed were deposited into an archive. Using this list, they can compare the collection in ACE with what they believe was deposited.

In addition to generating web page reports, all reports, status, item details, and event log queries can be exported in JSON (JavaScript Object Notation) format. This allows libraries to include integrity information from ACE in any collection portals they may develop.

4.3.5 Access Control
The Audit Manager provides access controls over to various functions of the web portal. This allows managers to create different usernames and passwords having different roles within the Audit Manager. For example, an account may be created for ‘browse’ level access which gives read-only access to collections and items in the collection. The browse account may be able to view log files, tokens, file information, but not able to remove items from monitoring or modify collection parameters. The following table shows which access controls are available.

<table>
<thead>
<tr>
<th>Access</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Collection Modify</td>
<td>Modify the connection parameters for a collection.</td>
</tr>
<tr>
<td>Browse</td>
<td>Browse files in a collection and display general metadata</td>
</tr>
<tr>
<td>Audit</td>
<td>Start a file or token audit on a collection</td>
</tr>
<tr>
<td>View Report</td>
<td>View a report showing missing or corrupt items</td>
</tr>
<tr>
<td>Remove Item</td>
<td>Remove an item from monitoring</td>
</tr>
<tr>
<td>Users</td>
<td>Add or modify users</td>
</tr>
<tr>
<td>Download Token</td>
<td>Down integrity tokens attached to files</td>
</tr>
<tr>
<td>Download Item</td>
<td>Download the monitored file</td>
</tr>
</tbody>
</table>

5. ACE Use Cases
ACE has been extensively tested, first in the Transcontinental Persistent Archival Prototype (TPAP) and second in the Chronopolis Project. These two testbeds tested ACE against individual collections several terabytes in size and containing several million files. In addition, the TPAP testbed also spanned three different storage types including iRODS[4], SRB[5], and
local file storage. Overviews of these collections are shown in table 2.

<table>
<thead>
<tr>
<th>Installation</th>
<th>Collections</th>
<th>Items</th>
<th>Log Events</th>
</tr>
</thead>
<tbody>
<tr>
<td>TPAP</td>
<td>32</td>
<td>1,505,392</td>
<td>3,030,152</td>
</tr>
<tr>
<td>Chronopolis</td>
<td>4</td>
<td>3,903,922</td>
<td>8,059,007</td>
</tr>
</tbody>
</table>

The TPAP installation of ACE is currently being used to monitor files in a variety of storage architectures at the University of Maryland and San Diego Supercomputing Center. The installation supports collections stored on both the SRB and iRODS. Data is a mix of small text files and large images.

The Chronopolis installation actually involves three independent Audit Monitors installed at the University of Maryland, San Diego Supercomputing Center, and the National Center for Atmospheric Research. Collections are stored in the SRB. Data from the three installations are aggregated into a common portal. This allows depositors to view the overall status of their collections easily without connecting to each site. In addition, the collection comparison functionality is used to ensure that identical copies of each collection exist at all sites.

Collections in Chronopolis vary in both total file count and average file size, allowing us to explore how file size affects collection auditing. The current archive policy is to audit files at UMIACS every 30 days. In Chronopolis, we discovered most of the delay in processing small files was due to SRB overhead, to prevent this from negatively impacting audit speed, ACE audited each collection using 5 threads reading files in parallel. This allowed metadata operations to run in parallel with data retrieval operations. The table below shows the results of these audits. While the SRB can sustain more simultaneous, no more than 5 threads were used to ensure other access to the archive was not impeded.

<table>
<thead>
<tr>
<th>Installation</th>
<th>Files</th>
<th>Directorie s</th>
<th>Size</th>
<th>Time(h)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CDL</td>
<td>46,762</td>
<td>28</td>
<td>4,291 Gb</td>
<td>20:32</td>
</tr>
<tr>
<td>SIO-GDC</td>
<td>197,718</td>
<td>5,230</td>
<td>815 Gb</td>
<td>6:49</td>
</tr>
<tr>
<td>ICPSR</td>
<td>4,830,625</td>
<td>95,580</td>
<td>6,957 Gb</td>
<td>122:48</td>
</tr>
<tr>
<td>NC-State</td>
<td>608,424</td>
<td>42,207</td>
<td>5,465 Gb</td>
<td>32:14</td>
</tr>
</tbody>
</table>

A usability test was performed over the summer of 2008 at the SAA Electronic Records Summer Camp. This test involved over 40 archivists and librarians from a variety of non-technical backgrounds. Participants were asked to use the Audit Manager to audit collections stored in the iRODS environment. Most participants were able to successfully audit their collections with less than two hours exposure to the technology.

6. CONCLUSION
In this paper we have described a software system that implements the ACE platform integrity management. We have described an Audit Manager component that is a low maintenance, highly scalable solution for archives and digital libraries to monitor the integrity of their digital assets.

7. REFERENCES
A Digital Library Service for the Small

Stavros Angelis
Dimitris Gavrilis
Panos Constantopoulos
Christos Papatheodorou

Digital Curation Unit, "Athena" Research Centre
Artemidos 6 & Epidavrou, Maroussi
Athens, Greece

\{ s.angelis, p.constantopoulos, d.gavrilis, c.papatheodorou \}@dcu.gr

ABSTRACT

In this paper, we present MOPSEUS, a lightweight digital library service based on the Fedora system. This service was created to address the needs of small libraries without support from technical staff. Hence, MOPSEUS attempts to balance flexibility against ease of installation, configuration and use. The service is available as a standard Java Web servlet, uses no external databases or other systems and can easily be deployed on top of any Fedora installation. Preliminary tests concerning the ease of installation and use are encouraging. We contend that facilitating the introduction of digital library infrastructures in the small may contribute to spreading digital curation practices.

Categories and Subject Descriptors
D.2.11 [Software Architectures]: Domain-specific architectures,

General Terms
Management, Documentation, Design, Experimentation, Human Factors, Standardization.

Keywords
Digital libraries, metadata schema, Fedora digital library, open architecture, preservation.

1. MOTIVATION

In the last decade, the library community, supported by the advances in information and communication technologies, demonstrated a rapid growth characterized by its entry into to the digital world. This move goes along with the emergence of a large number of complex digital library systems (both commercial and open source) that can handle millions of records and can serve large numbers of users with various characteristics and needs. However, the target group of such systems consists mostly of medium to large size libraries with sufficient budget and technical staff to acquire and maintain them. If the library community were sorted by size, the following categories would emerge:

- **Large size libraries**: usually national libraries or academic libraries.
- **Medium size libraries**: usually found in companies, small universities, branches of large academic libraries or technical schools.
- **Small size libraries**: usually found in schools and less populated areas, such as small towns and villages.

These three categories have different characteristics (such as budget, number of employees, and number of users) that can be seen in Table 1 below. The budget, employee and user numbers increase with library size, whereas the category population grows inversely to the library size.

<table>
<thead>
<tr>
<th>Categories</th>
<th>Large size</th>
<th>Medium size</th>
<th>Small size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number</td>
<td>small</td>
<td>small - medium</td>
<td>large</td>
</tr>
<tr>
<td>Budget</td>
<td>large</td>
<td>medium</td>
<td>small - very small</td>
</tr>
<tr>
<td>Number of employees</td>
<td>large</td>
<td>medium - small</td>
<td>small - very small</td>
</tr>
<tr>
<td>Number of users</td>
<td>large - very large</td>
<td>medium</td>
<td>small - very small</td>
</tr>
</tbody>
</table>

In Greece, for example, there are a few tens of large libraries (mostly academic) and a few hundred small size ones (mostly found in small islands or schools). Similar examples can be found all over the world [4]. The financial and technical requirements of a modern digital library system exclude a large community of libraries, namely the small ones [5]. The small libraries community usually has inadequate funds and staff to use such systems, even if these are free and open source [6].

The small library community cannot afford expensive commercial library systems, but also faces great difficulties in setting up and maintaining most open source systems because of lack of relevant expertise [7]. Small libraries usually have no more than two employees, who are often volunteers and in most cases non-technical. Therefore, we set forth the following set of requirements for a digital library service suitable for employment in small libraries:

- **Ease of installation**: someone without technical background should be able to install the digital library quickly and easily in a few steps.
• **Ease of configuration:** the librarian should be able to configure the system quickly and easily.

• **Flexibility:** the system must be able to support different, custom metadata schemata. This would enable the service to adapt to different library needs (e.g. school, museum library, public library, etc.).

• **Web-based:** users should be able to access the system remotely through the Internet and through a familiar overall interface. In addition, it should be possible to setup the system on a single server (e.g. a small library server/workstation) and provide access to more than one librarians/users.

• **Interoperability:** the system must be compatible with other systems and able to expose its collection to external agents (e.g. PMH harvesters).

• **Preservation and backup:** the system must be able to handle data preservation and allow the user to easily backup the database.

This paper presents MOPSEUS, a digital library service based on the idea to address the needs of this community for ease of installation, configuration and use. Furthermore, the initiative to facilitate the introduction of digital library infrastructures in the small libraries, may contribute to spreading digital curation practices. The availability of a simple and convenient tool like MOPSEUS to small library communities enables them to perform simple, yet crucial operations required for digital curation, such as the ingestion, modification and processing of digital objects, while enabling the specification of useful metadata, and providing a simple backup process. Furthermore, ongoing work on MOPSEUS focuses on providing mechanisms for the transformation and interoperability between different metadata schemata. The aforementioned features of MOPSEUS are still a long way from satisfying all the requirements for digital curation, as presented in the lifecycle model of the UK’s Digital Curation Centre (DCC) [9] or the more recent DCC&U lifecycle model [10] of the Digital Curation Unit (DCU). However, MOPSEUS still provides an initial subset of necessary operations for digital curation. These features may act as a stepping stone, making it far easier, thus acceptable, for small library communities to consider satisfying additional curation operations, as specified in [9,10].

2. **SYSTEM ARCHITECTURE**

In order to address the above requirements, two widely recognized open source systems were considered: DSpace [3] and Fedora [1]. The first, though generally adopted by the majority of the academic library community, has a more complex installation and configuration process. Furthermore, its metadata schema and preservation architecture do not meet the flexibility requirement. On the other hand, the Fedora system has an open, service oriented architecture and an easy deployment process, thus better meeting most of the requirements:

• **Interoperability:** Fedora’s service oriented architecture and embedded OAI provider greatly enhance its interoperability features [2].

• **Preservation & backup:** Fedora stores its digital objects in the file system and a simple backup process is reduced to just copying a specific folder to some external media.

• **Flexibility:** Fedora’s object architecture involves multiple datastreams and an interconnection capability through the Fedora ontology, thus making Fedora extremely flexible.

The main drawback of Fedora, on the other hand, is its lack of a user-friendly interface.

2.1 **MOPSEUS**

MOPSEUS is a Web based system built on top of Fedora and consists of three main subsystems (Figure 1): communication, datastream management, user functions. The communication subsystem employs the SOAP and REST protocols in order to communicate with the Fedora server.

![Figure 1. The MOPSEUS Architecture](image)

The user interacts with the Fedora system through a set of basic functions that include: object ingest, modify, purge, object relations. These functions refer to Fedora digital objects that contain datastreams. These datastreams are handled by a separate subsystem that is responsible for rendering the datastream contents to the user through a friendly interface. Furthermore, the user can extend the system by adding new datastreams.

2.2 **Metadata schema**

In contrast to most Fedora-based systems, MOPSEUS does not require an external database management system (such as MySQL or Oracle), but rather uses Fedora’s internal McKoi database system instead. This approach greatly simplifies the installation process since the user only has to install Fedora (a very simple process) and then simply copy the MOPSEUS .war file to the Fedora installation. The problem with not having an external database system is where to store all the configuration data that a digital library needs. The solution is to encode the configuration data as digital objects in the repository itself, thus creating a self-describing repository. A separate namespace (admin) was created to hold these data.

The metadata schemata available in MOPSEUS are described in XML and stored as datastreams in a specific digital object. For
each entry in the metadata schema list a separate datastream must be created in the digital object in order to describe that schema. Only the Dublin Core metadata schema is mandatory, a choice intended to ensure interoperability. For example, if the user wants to extend the existing metadata schemata by including a new one (e.g. MODS), he must add an entry to the list of the available metadata (e.g. with ID=MODS) schemata and then create a new datastream with label MODS that describes the MODS metadata schema using XML. An example of this datastream can be seen in Figure 4. MOPSEUS will translate this XML based schema into user-friendly HTML forms.

```xml
<schema id="MODS" name="MODS Schema">
  <element id="title" name="title" type="text">
    <format attribute="false" attribute_name="" write="false" write_id="true" /></format>
    <style css="normal_text" size="40" type="text">
      <display>Title</display>
      <description>Enter the title of the article</description>
    </element>
  <element id="type" name="type" type="list">
    <format attribute="true" attribute_name="type" write="true" write_id="true" /></format>
    <style css="list_text" size="40" type="select">
      <display>Publication Type</display>
      <description>Enter the publication type of the article</description>
      <value>Journal</value>
      <value>Conference</value>
      <value>Book</value>
    </element>
</schema>
```

**Figure 2. An Example of a Metadata Schema Description**

In Figure 2, we give an example of a metadata schema description in XML, namely a sample MODS metadata schema containing two elements, title and type. The user may define an arbitrary number of elements and nest them in every possible way. For each element, the user has to provide the following information:

- **Basic information**: element’s id, name and type. Type can have the following values: label, text, list. Label is used to group elements together. Text is used to present the user with various text fields. List provides the user with select lists.
- **Format information**: the format information describes how the element data will be encoded when creating the XML datastream. For example, attribute, write, write_id.
- **Style information**: the style information enables the user to choose the style class that will be used for rendering the element in html along with its size (e.g. in the case of text area box: rows and columns).
- **Display and description information**: the display and description information defines the label and a short description of its function.

### 2.3 Web based interface

In order to overcome Fedora’s lack of a user friendly interface, a Web-based interface was built on MOPSEUS. The embedded tomat Web server that came with Fedora made deployment even easier since the user only needs to copy the MOPSEUS .war file to a specific directory of the tomat server.

Basic functions supported include search, ingest, modify and purge.

![Figure 3. The Basic Search Screen](image)

The main page of MOPSEUS (SEUS) displays the digital objects contained in the library, allows the user to ingest an object and perform basic search functions. These functionalities are provided by a user-friendly web-based interface, minimizing the technical skills required by the cataloguer.

By clicking on an object, the user is presented with its Dublin Core record and a list of the other available datastreams of that object (Figure 4). The user can also insert a new datastream, or purge an existing one. The fact that most of the functions of MOPSEUS are handled from these two pages minimizes the learning overhead.

### 3. PRELIMINARY EVALUATION

A preliminary experimental investigation of the MOPSEUS system usability has been performed using an expert – based methodology [8]. In particular the Cognitive Walkthrough method was selected which concentrates mainly on the difficulties users may experience in learning to operate an application to perform a given task. During a Cognitive Walkthrough process the expert first determines the exact sequence of correct task performance, and then estimates, on a screen by screen basis, the likely success or failure of the user in performing such a sequence. The main characteristic of this approach is that the expert must make an informed guess of the likely reaction of users and explain why certain interface attributes are likely to cause users difficulties.

172
In our experiment ten librarians with significant expertise in the development of digital collections and the usage of digital library technologies were asked to install the system, and use it by ingesting and retrieving digital objects and metadata. After a 5-days usage period the participants were asked to fill a short questionnaire, expressing their attitude on a 5-point Likert scale towards a set of criteria including learnability, ease of use, navigation, aesthetic appearance, use of established terminology and finally the perceived usefulness of the system’s functions (i.e. their perception of the improvement of their productivity on digital collections management).

To better assess whether the system fits and supports adequately their everyday tasks, a part of the questionnaire focused on particular functions, such as searching and browsing, ingesting

<table>
<thead>
<tr>
<th>Item</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Learnability</td>
<td>4.2</td>
</tr>
<tr>
<td>Aesthetic appearance</td>
<td>3.8</td>
</tr>
</tbody>
</table>

Furthermore, some users requested an advanced search facility and the ability to create collections and add items to them. The lack of these functionalities is responsible for the low usefulness average. Regarding the collection definition and development functionality, it should be mentioned that the users were not familiar with the Fedora relationships that allow the user to create collections of items. Regarding the installation process, a simpler installation procedure was created in response that helped users perform the installation quicker and easier. Finally, it should also be noted that the aesthetics evaluation which rated relatively low, had a large standard deviation (about 1.5) while the other evaluation criteria, had a standard deviation of approximately 0.5.

4. CONCLUSIONS AND FUTURE WORK

MOPSEUS is a lightweight, open SOA digital library service based on Fedora, easy to install and maintain, provided as open source by the Digital Curation Unit. These features make MOPSEUS suitable for small libraries with no specialized technical staff. At this early stage MOPSEUS has shown some promising results. Several more features need to be implemented, such as workflow management, advanced search and easier parameter configuration.

5. ACKNOWLEDGMENTS

We would like to thank Prof. Antonios Deligiannakis for reviewing this paper.

6. REFERENCES


http://www.dlib.org/dlib/april03/staples/04staples.html


http://smallschools.ischool.washington.edu/

http://www.linccweb.org
[6] Calhoun, K., Hello and On Small Libraries,
http://community.oclc.org/metalogue/archives/2008/06/hello-and-on-small-libraries.html

[7] Fuller, D., School Library Journal & San Jose State
University 2006 Automation Survey: The systems are
changing. But school libraries aren’t. School Library Journal
10-01-2006
http://www.schoollibraryjournal.com/article/CA6376081.html

[8] Nielsen, J., Usability engineering. Morgan
Kaufmann, Boston, USA (1993).

Proceedings of the 8th ACM/IEEE-CS joint conference on
Digital libraries. Pittsburgh PA, PA, USA, pp. 453-453.

[10] Constantopoulos, P., Dallas, C., Androutsopoulos, I.,
Angelis, S., Deligiannakis, A., Gavrilis, D., Kotidis, Y.,
Papatheodorou, C., 2008. DCC&U: An Extended Digital
Curation Lifecycle Model. Proceedings of the 4th
International Digital Curation Conference. Edinburgh,
Invited Paper: Lessons Learned from the DISC-UK DataShare and Data Audit Framework Implementation Projects

Robin Rice
University of Edinburgh
EDINA and Data Library
Edinburgh, Scotland
R.Rice@ed.ac.uk

ABSTRACT
This paper discusses some of the outcomes from two JISC-funded data curation projects in the UK from 2007-2009: DISC-UK DataShare and the Edinburgh Data Audit Framework Implementation project. DISC-UK DataShare involved investigating deposit of research data in institutional repositories including metadata solutions and policy development; the second was about understanding and improving data management practice through partnering with academic departments in the use of the Data Audit Framework.

1. ABOUT THE DISC-UK DATASHARE PROJECT
DISC-UK (Data Information Specialists Committee - United Kingdom) is a forum for data professionals working in UK Higher Education who specialise in supporting their institution's staff and students in the use of data for analysis (primarily statistical and geo-spatial). This partnership, led by EDINA and the Data Library at the University of Edinburgh, carried out the DISC-UK DataShare project (March 2007 - March 2009): http://www.disc-uk.org/datashare.html. The aim was to explore new pathways to assist academics wishing to share their data over the Internet. With three institutions taking part – the Universities of Edinburgh, Oxford and Southampton – plus the London School of Economics as an associate partner, a range of exemplars have emerged from the establishment of institutional data repositories and related services. It was part of a wider programme to develop institutional repositories funded by the UK’s Joint Information Systems Committee (JISC).

Part of the variety in the exemplars was due to the different repository platforms used by the three project partners: DSpace, ePrints and Fedora (all open source software). Another was due to different approaches in setting up the repositories. All three institutions had an existing, well-used institutional repository, but two chose to incorporate datasets within the same system as the publications, and one (Edinburgh DataShare) was a paired repository exclusively for datasets, designed to interoperate with the publications repository (Edinburgh Research Archive).

2. TOPICS EXPLORED
A number of scoping activities were carried about by the partners with the goal of informing repository enhancement as well as broader dissemination. These included a State-of-the-Art-Review to determine what had been learned by previous repository projects in the UK that had forayed into the data arena. This resulted in a list of benefits and barriers to deposit of datasets by researchers to inform our outreach activities [1]. A Data Sharing Continuum was developed to illustrate where the projects were aiming to fit into the curation landscape, and the range of curation steps that could be taken, from simple backup to online visualization [2]. Later on, a specialized metadata schema was explored (Data Documentation Initiative or DDI) in terms of how it might be incorporated into repository systems, though repository development in this area was not taken up [3]. Instead, a dataset application profile was developed based on qualified Dublin Core (dcterms). This was implemented in the Edinburgh DataShare repository and adapted by Southampton for their next release [4]. The project wished to explore wider issues with open data and web publishing, and therefore produced two briefing papers to do with data mashups – on numeric data and geospatial data [5, 6]. Finally, the project distilled what it had learned in terms of policy development for data repositories in a training guide [7]. A fuller discussion of the project as a whole is available [8].

3. DATA AUDIT FRAMEWORK IMPLEMENTATION
The Data Audit Framework (DAF) Implementation projects – along with the Data Audit Framework Development project (DAFD) led by DCC/HATII - were conceived in response to recommendations made by Liz Lyon in the seminal JISC-commissioned report Dealing with Data: “A framework must be conceived to enable all universities and colleges to carry out an audit of departmental data collections, awareness, policies and practice for data curation and preservation.” [9] The DAFD project developed the methodology starting in April, 2008, and the Edinburgh project started in May 2008, as one of four JISC-funded projects to test the framework through implementation. The Edinburgh project finished end of November, 2008.

The aim of the project was to provide exemplars of Data Audit Framework adapted to the current needs of data curation activities.
in University of Edinburgh, and to find what data the research community has, where it is located, and who is (or is not) responsible for it.

In order to be able to clearly address the issues in managing research data assets within the University we set out to implement the Data Audit Framework across three colleges, therefore a range of disciplines.

Our overall approach consisted of an online questionnaire and semi-structured interviews with research active staff across three colleges. In terms of the methodology we used the spreadsheet-based Data Audit Framework methodology developed by the DAFD project team in advance of the online tool. The methodology was implemented in four stages:

1. Planning the audit
2. Identifying and classifying assets
3. Assessing management of data assets
4. Reporting and recommendations

The key outputs from the project were five case studies – the final ‘audit’ reports for each research unit audited [10]. These were the School of Divinity, Economic and Social History, Centre for Integrative Physiology, Brain Imaging, and the Institute for Astronomy. None of the audits was a comprehensive survey of research data but they proved to be a good starting point to auditing research data holdings and investigating data management practices within the University.

A full description of the Data Audit Framework and lessons learned from the implementation projects is available [11].

4. REFERENCES
Preservation Workflows, Strategies and Infrastructure

David Giaretta
(on behalf of the CASPAR project)
Science and Technology Facilities Council
Rutherford Appleton Lab
Didcot, Oxon, UK
david.giaretta@stfc.ac.uk

ABSTRACT
The OAIS Reference model [1] provides the de-facto model for digital archives and forms the basis for the effort to produce an international standard for audit and certification of digital repositories. OAIS is however an abstract model; digital repositories must have concrete tools, strategies and appropriate support to implement the requirements which derive form OAIS.

This paper introduces and briefly describes a number of fundamental strategies, workflows, and a support infrastructure to enable repositories to follow OAIS and to help them be better positioned for international certification.

Categories and Subject Descriptors
H1.1 [Models and principles]: Systems and Information Theory; H3 [Information storage and retrieval]; K.6.1 [Management of Computing and Information Systems]: Strategic information systems planning

Keywords
digital preservation, digital curation, infrastructure, workflows.

1. INTRODUCTION
The OAIS Reference Model provides a number of models for repositories including a Functional Model, to which is relatively easy to map an existing archive system, an Information Model, which is rather more challenging, an Information Packaging Model and federation models, plus preservation perspectives including types of migration and a variety of software related processes. A number of overall strategies, processes and supporting infrastructures may be derived from these.

The mandatory responsibilities that an organization must discharge in order to operate an OAIS archive are:

- Negotiate for and accept appropriate information from information Producers.
- Obtain sufficient control of the information provided to the level needed to ensure Long-Term Preservation.
- Determine, either by itself or in conjunction with other parties, which communities should become the Designated Community and, therefore, should be able to understand the information provided.
- Ensure that the information to be preserved is Independently Understandable to the Designated Community. In other words, the community should be able to understand the information without needing the assistance of the experts who produced the information.
- Follow documented policies and procedures which ensure that the information is preserved against all reasonable contingencies, and which enable the information to be disseminated as authenticated copies of the original, or as traceable to the original.
- Make the preserved information available to the Designated Community.

This paper describes some of the strategies, workflows and infrastructure components which we argue follow from OAIS and which are being pursued by the CASPAR project (CASPAR 2007 [2] and the UK Digital Curation Centre [6]) concludes with a description of the status of the work on audit and certification of digital repositories and a brief description of an infrastructure to support preservation.

2. CASPAR/DCC APPROACH
Regarding the OAIS mandatory responsibilities the CASPAR project (CASPAR 2007 [2]) argues that one can derive a number of requirements for information flow. For example an archive needs to identify/capture

1) the access and digital rights management (DRM) associated with a digital object
2) the Representation Information – shown in Figure 1
3) the Designated Community
4) an appropriate storage mechanisms
5) an infrastructure to support preservation over the long term

In addition, the artifacts, such as the DRM and Representation Information, must themselves be preserved over the long term.

Further details of CASPAR are available from the CASPAR web site http://www.casparpreserves.eu and the DCC web site.
There are a number of ways in which, for example, Representation Information, may be captured, and a number of strategies must be examined. The following section illustrates some of these strategies.

The model discusses the different levels of application of OAIS. It identifies what components of a CASPAR-based archival system are required, their functions and how they must interoperate at a high level. These components include for example a Representation Information Toolbox, a Virtualisation Assistant and a Preservation Orchestration Manager. The Conceptual Model also surveys the key underlying technologies of knowledge management and virtualisation that underpin the entire CASPAR approach. It touches on all aspects of the CASPAR approach to the preservation of digitally encoded information, and is reflected in the CASPAR architecture.

4. PRESERVATION STRATEGIES
It is sometimes argued [4] that the two preservation strategies available are emulation and migration. In fact there are a number of strategies which may be adopted, each having its limitations as discussed next.

4.1 Emulation
Emulation has been defined as “the ability of a computer program or electronic device to imitate another program or device.” (Wikipedia, emulation [5]). The fundamental aim is to do with current hardware and software what could be done in the past. This is adequate for rendering documents and running applications such as computer games on single machines. However it is very limiting in the context of data because at least sometimes (and perhaps most often) one does not simply reproduce what has been done previously; rather one wants to re-purpose and re-analyse archived data using modern tools and techniques. In addition emulation currently has difficulties with network aware applications and processes.

It is worth noting that emulation systems may be regarded as special types of Representation Information in that it may assist the understanding of digitally encoded information.

4.2 Access Software
Access Software (OAIS 2002) “presents some or all of the information content of an Information Object in forms understandable to humans or systems. It may also provide some types of access service, such as displaying, manipulating, processing, or sub-setting, to an Information Object.”

This allows one to “plug-in” to new software to access information encoded in digital objects. However there may be a mismatch in concepts. CASPAR addresses this through the application of virtualisation techniques which attempt to identify common and discipline related types of objects, for example images or tables, on the assumption that future systems will employ these concepts and so the “plug-in” is likely to be more easily and reliably developed.

4.3 Migration
The primary types of migration, ordered by increasing risk of information loss, are (OAIS 2002):

- Refreshment: A Digital Migration where a media instance, holding one or more AIPs or parts of AIPs, is replaced by a media instance of the same type by copying the bits on the medium used to hold AIPs and to manage and access the
medium. As a result, the existing Archival Storage mapping infrastructure, without alteration, is able to continue to locate and access the AIP.

- Replication: A Digital Migration where there is no change to the Packaging Information, the Content Information and the PDI. The bits used to convey these information objects are preserved in the transfer to the same or new media-type instance. Note that Refreshment is also a Replication, but Replication may require changes to the Archival Storage mapping infrastructure.

- Repackaging: A Digital Migration where there is some change in the bits of the Packaging Information.

- Transformation: A Digital Migration where there is some change in the Content Information or PDI bits while attempting to preserve the full information content. It is this type of migration which is usually referred to in the context of “emulate or migrate”.

One important consideration is the need to analyse the implications of any type of migration, and in particular for repackaging or transformation. Regarding transformation one needs to think carefully, for example about:

- the potential loss of information – especially if special conventions have been adopted on top of particular formats, which are then embodied in access software. In addition the underlying information concepts which are captured in the initial and final forms may not match – here the virtualisation concepts from CASPAR can help.

- the associated costs of transformation, for example whether it should be done in bulk or on-demand.

- the implications for authenticity – how can one prove that the transformed version is indeed the “same” as the original

- the preservability of the new form

### 4.4 Description Techniques

Representation Information may also include the description of the structure and the semantics of the digitally encoded object. CASPAR is developing and bringing together many techniques for producing and validating this type of description.

The question of how much Representation Information and whether it is adequate is addressed in OAIS through the concept of Designated Community. CASPAR is demonstrating techniques for validating the types and quantity of Representation Information by parsing the data using the descriptions, analogous to the way in which XML is validated. In addition we use the descriptions in generic applications to show to the satisfaction of the data experts that one can process and analyse the data object and produce the same results as with the software normally associated with it. Note that these generic applications are not meant as to replacement data specific current applications not least because the generic applications are slower and have limited functionality.

### 5. STRATEGY SELECTION

Figure 2 shows the workflow appropriate for selecting preservation strategies which includes analyses of the archive, preservation objectives and, very importantly, identification of the Designated Community. Each strategy is evaluated in a cost/benefit analysis which is, it must be admitted, still quite rudimentary.

![Figure 2 Preservation Strategy selection workflow](image)

An example of strategy evaluation may be shown from CASPAR where the issue is the continued use of the proprietary MAX/MSP software which runs on Macintosh computers.

<table>
<thead>
<tr>
<th>Details/Strategy and related scenarios</th>
<th>Difficulties</th>
<th>Expected outcomes</th>
<th>Evaluation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Emulation on SUN or LINUX</td>
<td>Emulation on windows is possible using existing commercial and public domain software. However the timing of the execution would be an issue. Also drivers for any special speaker hardware may be needed for specific pieces. Simple pieces such as XXXX and XXX should be no problem but YYYY and ZZZZ would need to be investigated.</td>
<td>The MAX/MSP software would run but performance and driving modern hardware may be a problem.</td>
<td>Performance needs to be investigated.</td>
</tr>
<tr>
<td>Use alternative PD software</td>
<td>Public domain software is available - see Wikipedia about MAX/MSP, however compatibility issues must be considered.</td>
<td>- Some basic (primitive) operations are not available in PD. - Max/MSP patches are frequently dependent on « externals</td>
<td>Solutions of this kind already</td>
</tr>
</tbody>
</table>
investigated. » - It is possible to write externals and patches that work with Max/MSP and Pd using flexl and cyclone. explored

Get code from commercial providers (to be held in escrow) This depends upon our relationship with the company. One strong argument we can use is that if they do not do this then people will stop using their software

Re-implement workflow that MAX/MSP does Needs detailed analysis of the workflow activities and also the individual signal processing modules that the workflow links together

Potentially a lot of work but could be worthwhile if we have enough effort and if we write the software in a preservable way.

Depends on contacts with commercial suppliers

Cost could be high but potentially very valuable

6. PRESERVATION INFRASTRUCTURE AND PRESERVATION WORKFLOWS
A number of other workflows arise from the support components identified by CASPAR, which may be summarized in Figure 3.

Figure 3 contains a number of information flows; some sequences of these flows making up workflows important for digital preservation and two of these are described next.

6.1 Workflows for use of digital objects
The following workflow, extracted from Figure 3, illustrates the way in which digital objects may be used and understood by users.

Identifiers (called here Curation Persistent Identifiers - CPID) are associated with any data object, which point to the appropriate Representation Information in a Registry/Repository, as illustrated in Figure 4. The Representation Information returned by the Registry/Repository itself is a digital object with its own CPID.

Figure 3 Preservation Workflows

The above is not meant to imply that there must be a single, unique, Registry/Repository, nor even a single definitive piece of Representation Information for any particular piece of digitally encoded information.

6.2 Workflows for maintaining the Representation Network
The Registry/Repository is supplemented by the Knowledge Manager – more specifically a Representation Information Gap manager which identifies gaps which need to be filled, based on information supplied to the Orchestration component.

Of course the information on which this is based does not come out of thin air. People (initially) must provide this information and the Orchestration Manager collects this information and distributes.
Support for automation in identifying such “gaps”, based on information received, is illustrated in Figure 5 which shows users (u1, u2...) with user profiles (p1, p2... – each a description of the user’s Knowledge Base) with Representation Information {m1, m2,...} to understand various digital objects (o1, o2...).

Take for example user u1 trying to understand digital object o1. To understand o1, Representation Information m1 is needed. The profile p1 shows that user u1 understands m1 (and therefore its dependencies m2, m3 and m4) and therefore has enough Representation Information to understand o1.

When user u2 tries to understand o2 we see that o2 needs Representation Information m3 and m4. Profile p2 shows that u2 understands m2 (and therefore m3), however there is a gap, namely m4 which is required for u2 to understand o2.

For u2 to understand o1, we can see that Representation Information m1 and m4 need to be supplied.

This illustrates one of the areas in which Knowledge Management techniques are being applied within CASPAR, in addition to the capture of Semantic Representation Information.

7. AUTHENTICITY WORKFLOW

7.1 Authenticity Protocol (AP)
The protection of authenticity and its assessment is a process. In order to manage this process, we need to define the procedures to be followed to assess the authenticity of specific type of objects.

7.2 Overall Authenticity Model
The overall model is designed as shown in the diagram below:

8. DIGITAL RIGHTS PROCESSES AND WORKFLOWS
The processes and workflows associated with the capture and preservation of Digital Rights of a digital object may be illustrated as follows:

8.1 Request of a license on a digital object
Here is a sequence diagram describing the rights distribution, verification and enforcing process. The information flow for acquiring a digital right on a digital content item is the following:

1. A user selects a content object he has found searching in the CASPAR archive. If the content is protected by digital rights and the rights holder was willing to give a usage licence, the user is presented a list of license offers. He chooses one amongst them and a new user license is generated.

2. The user requests to access the content object by invoking a service. Before the service can be invoked, a security check has to be performed by the Data Access Manager and Security (DAMS) module. One of the tasks of this security check is to ask the DRM layer whether the content is protected by digital rights and, in case of a positive answer, if the user owns a valid license. If this verification is successful the DAMS grants access to the resource letting the service to be executed.
8.2 Modification in the legal framework

Here is a sequence diagram describing the preservation of the license offers after a modification in the legal framework.

![Sequence diagram](image)

Figure 8 DRM preservation

Depending on the particular change in the Legal Framework, the above sequence diagram could have a variant, i.e. concluding with the update of a Legal Framework Context object, instead of the update of some license offers.

9. WORKFLOWS FOR AUDIT AND CERTIFICATION OF DIGITAL REPOSITORIES

CASPAR aims to research, implement and put into practice innovative solutions for preservation of digital resources. Such resources will be held over the long term in a variety of repositories, operated by different organisations, funded and managed in different ways, and with different policies and procedures to safeguard their contents. How can we be confident that a repository, even if it is based on the CASPAR methods and tools can be trusted over long timescales to preserve the resources with which it is entrusted? This question is the motivation for auditing and certifying digital repositories in a systematic way. Repositories might wish to audit themselves to identify any weaknesses in their procedures or they might prefer to be formally audited by an independent body to provide assurance to their funders or users of their trustworthiness.

There have already been a number of initiatives towards audit and certification of digital repositories. In 2003, the RLG (now part of OCLC) and NARA created a joint task force which produced the influential document “Trustworthy Repositories Audit and Certification: Criteria and Checklist”. This document defines three high-level areas in which a repository must be trustworthy: organisational infrastructure, digital object management and technologies, technical infrastructure and security. Within these areas, a number of criteria are defined that a repository should satisfy, and examples are given of the kind of evidence that would show this. Also worthy of mention is the work of nestor (Network of Expertise in Long-Term Storage of Digital Resources) in Germany which has designed a catalogue of criteria for trusted digital repositories for long-term preservation, and DRAMBORA (Digital Repository Audit Method Based on Risk Assessment) developed jointly by the Digital Curation Centre (DCC) and DigitalPreservationEurope (DPE).

Work is going on towards a full ISO standard for audit and certification; it will be through the ISO TC20/SC3 committee. The CASPAR project is contributing to this effort. Regular ‘virtual meetings’ are taking place, plus some face-to-face meetings, at which the above-mentioned documents and approaches are being re-examined and adapted into a form suitable for direct use in audit and certification, with an emphasis on clarity and usability. There is a public website at http://wiki.digitalrepositoryauditandcertification.org/ at which the progress of the work can be seen. It is possible to request registration so as to be able to edit the content of the site (which is a collaboratively edited wiki) and take part in the meetings.

The workflow arising from Audit and Certification includes:

- creation of the appropriate international (ISO) standards
- creation of the accreditation body
- accreditation of the certification organisations

The audit and certification of individual archives will involve a tailored workflow which includes gathering the appropriate evidence, adjusting internal processes if needed, performing internal and trial audits, and finally undergoing the full audit. Regular repeats of the audits will of course be required.

10. PRESERVATION INFRASTRUCTURES
Figure 8 shows a possible preservation infrastructure, in the context of a broader e-Research infrastructure.

Analogy is drawn with the network infrastructure which links islands of connectivity (e.g. the EU funding of GEANT). The network supplies services to isolated organisations, each with its own resources such as CPU and storage. Additional infrastructure such as the EU’s EGEE project, and the UK’s various GRID project, connect these. Similarly infrastructure components connect the repositories which use the lower levels of functionality.

In the future there should be some similar interconnected repositories and we need to transfer information from the “now” to the future However there are many difficulties/threats to this transfer, listed below. To counter these threats we need to put in place components to support preservation.

The threats to digital preservation include:

- users may be unable to understand or use the data e.g. the semantics, format, processes or algorithms involved
- non-maintainability of essential hardware, software or support environment may make the information inaccessible
- the chain of evidence may be lost and there may be lack of certainty of provenance or authenticity
- access and use restrictions may fail in the future
- loss of ability to identify the location of data
- the current custodian of the data, whether an organisation or project, may cease to exist at some point in the future
- the ones we trust to look after the digital holdings may let us down

Many, if not all, of this preservation infrastructure is also useful to support the current e-Research infrastructure, particularly where in multidisciplinary research one needs to be able to use unfamiliar information resources.

Infrastructure components suggested here include:

- persistent identifier name resolver
- registry of representation information
- authenticity support tools
- processing context tools
- certification process
- orchestration/brokering
- knowledge gap manager

This is an area which needs further investigation and planning, as is being done by the PARSE.Insight project ([7])

11. CONCLUSIONS
Digital preservation is not a one-off activity; it involves a number of workflows which have been outlined in this paper. Further detail may be found at http://www.casparpreserves.eu; full documentation of the deliverables may be downloaded at http://www.casparpreserves.eu/publications/deliverables  The software may be found at http://developers.casparpreserves.
Many strategies must be considered and appropriate infrastructures are needed to share the effort of preservation.

12. REFERENCES


A Residential Data Curation Internship: Opportunities and Challenges

Kelly M. Gordon 1
Albert R. Mann Library
Cornell University
Ithaca, NY 14853 USA
kgordon@library.berkeley.edu

ABSTRACT
During the summer of 2008, while a student at San Jose State University’s School of Library and Information Science, the author completed a residential data curation internship at the Cornell Biological Field Station. The internship entailed preparing and documenting a long term dataset collected by researchers at the field station. Internships such as this one present learning opportunities not only for students interested in the field of data curation, but for researchers as well. They also provide the opportunity to facilitate collaborations between library staff and researchers.

1. BACKGROUND
Recognizing researchers’ need for data curation services early in the research cycle, librarians at Albert R. Mann Library at Cornell University applied for and received funding from the National Science Foundation to develop a local, institution-based data staging repository (DataStaR) for researchers affiliated with Cornell [1]. The intent of DataStaR is to provide researchers with the tools and support they need to perform data curation tasks in preparation for transmission of datasets to appropriate domain-based long-term data repositories. One goal of the project has been to foster partnerships with data owners in order to facilitate the development of this model.

One such partner is the Cornell Biological Field Station (CBFS), which serves as a primary field site for aquatic research at Cornell University. The centerpiece of the station’s research program is a 50-year dataset on the food web of Oneida Lake, New York. Researchers at the field station, increasingly aware of the need to document and preserve this resource, saw the potential benefits of participation in the DataStaR project. Since the DataStaR platform and tools were not yet ready for use by researchers, DataStaR and CBFS personnel decided to hire an intern to carry out the work of preparing the dataset and creating metadata for transmission to a data repository.

This work is licensed under the Creative Commons Attribution-Noncommercial-No Derivative Works 3.0 Unported license. You are free to share this work (copy, distribute and transmit) under the following conditions: attribution, non-commercial, and no derivative works. To view a copy of this license, visit http://creativecommons.org/licenses/by-nc-nd/3.0/.

2. INTERNSHIP
Mann Library and CBFS personnel collaborated in the hiring of an intern to complete a ten-week data curation internship during the summer of 2008. The successful candidate (the author of this paper) is an MLIS student enrolled in San Jose State University’s School of Library and Information Science. Since the intern would need to work closely with CBFS researchers to prepare the dataset and develop metadata, it was agreed that she would live and work at the field station, but be remotely supervised by Gail Steinhart, one of the primary investigators on the DataStaR project. CBFS hosts a field research internship program for undergraduates, so the data curation intern was one of eight interns working at the field station that summer.

The long-term database used by researchers at CBFS consists primarily of multiple tables in Microsoft Access format. The intern prepared data and created metadata for seven data packages derived from this dataset. Metadata were created using the EML (Ecological Metadata Language) metadata standard [2]. Since the DataStaR platform is still a work in progress, the intern used the software package Morpho [3] to create EML records. The data packages were deposited in the Knowledge Network for Biocomplexity [4], one of the pre-eminent domain-based repositories for ecological datasets, and Cornell University’s e-Commons [5], an institutional repository that houses electronic content produced by members of the Cornell Community.

3. OPPORTUNITIES
Residence at CBFS provided a number of opportunities for interactions above and beyond those related to the tasks of the internship. The intern’s involvement with the day-to-day activities of the field station provided a deeper understanding of the issues faced by researchers prioritizing data management tasks in the context of other activities in the research cycle. The intern also participated in field data collection and discussions of ongoing research. One-on-one interactions with field station personnel in the course of carrying out internship tasks provided insight into the varying attitudes of researchers towards data management activities.

The presence of an intern actively involved in data curation tasks presented learning opportunities for station personnel, as well. Although most experienced researchers are well aware of the
importance of data curation in a general sense and are eager to work to preserve and document their data, many lack an understanding of the nuts and bolts of data curation, and may not be aware of the role data repositories can play in data preservation. As the intern consulted with station personnel at each step of the process, researchers gained a deeper understanding of the tasks involved in the preparation of data and metadata for transmission to a data repository.

Students, new to research, may not have previously considered the importance of data management, especially given the greater attention that data collection and analysis are likely to receive from their advisors. Living and working with graduate and undergraduate students at CBFS provided the opportunity for the data curation intern to engage in a number of informal dialogs about the potential pitfalls of data neglect, and about the benefits of adequate data documentation and preservation.

4. CHALLENGES

Data curation tasks are likely to be pushed to the back burner by researchers engrossed by managing data collection efforts, analyzing data, and writing publications. One challenge encountered during this internship was the need to frequently consult with field station personnel who were often at conferences or collecting data in the field. Researchers at the field station were sensitive to the fact that their input was needed to keep the project moving forward, but some aspects of the work were occasionally stalled by the unavailability of one or more key people. This issue was ameliorated by communication regarding researchers’ and library staff schedules, and careful planning about the timeline on which specific tasks would be completed.

It is not uncommon for unanticipated issues to become evident when working with complex datasets that are handled by a number of different workers with varying degrees of familiarity with good data practices. Although some data preparation and cleanup tasks were planned for at the outset of this internship, it was necessary to assess each new issue as it arose to determine whether resolving it fell within the scope of the internship duties and whether it was feasible to attempt to correct the problem or if that responsibility would fall to the researcher. In general, when library personnel collaborate with researchers to perform data curation tasks, it may be desirable to define many data cleanup tasks as the responsibility of the researcher, in order to avoid an unsustainable drain on library personnel’s time.

Even the most user-friendly, flexible platform for data curation tasks may present barriers to use if the platform can’t be integrated into researchers’ existing workflows [6]. Researchers adopting new data curation practices may find that the learning and use of tools require unanticipated changes to their current data habits. Additionally, if data management tasks have been a lower priority in the past, datasets may be in need of more extensive preparation than is feasible for researchers. The need to prepare data may present the most daunting barrier to preservation of data, yet librarians can provide limited assistance with these sorts of tasks, since the specifics are peculiar to each data set.

The presence of an intern working on data curation tasks acts as a spur to researchers to provide necessary information and support to keep the work moving forward. However, the flip side of this is that when the internship ends, data curation tasks may be deprioritized. Additionally, the work done during this particular internship represented an intensive level of data service that, although eagerly accepted by CBFS researchers, is ultimately impossible for library personnel to sustain. Nevertheless, CBFS researchers have gained some familiarity with the tasks involved in preparing the data sets themselves for publication, and have agreed to assume responsibility for this portion of the process, while Mann Library staff continue to provide support for the creation of metadata. As the DataStaR platform is more fully developed and tools become available, CBFS personnel will assume greater responsibility for the entire publication process, with the librarian serving in an advisory capacity.

5. CONCLUSIONS

Graduate student internships can provide an excellent vehicle for the promotion of collaboration between researchers and librarians in accomplishing data curation goals. In particular, internships that embed the intern in the daily working environment of the researcher not only provide vital opportunities for interaction related to the task at hand, but also allow both researchers and the intern to benefit from an enhanced understanding of the particulars of data curation tasks, on the one hand, and of the research environment, on the other. Excellent communication regarding researchers’ schedules and the data services that will and will not be provided by the intern and other library personnel can help to smooth over some of the challenges that may be encountered during a collaboration such as this one.

6. ACKNOWLEDGMENTS

Sincere thanks to Gail Steinhart of Cornell University’s Albert R. Mann Library, and to Ed Mills, Lars Rudstam, Kristen Holecck, Tom Brooking, and the staff of Cornell Biological Field Station for their invaluable assistance. This work is supported by National Science Foundation grant number III-0712989.

7. REFERENCES


Preserving Electronic Mailing Lists: The H-Net Archive

Lisa M. Schmidt
Electronic Records Archivist, MATRIX
Michigan State University
East Lansing, MI 48823 USA
lisa.schmidt@matrix.msu.edu

ABSTRACT
This poster illustrates an NHPRC-funded project to evaluate and improve upon preservation practices for the H-Net academic e-mail list archive.

Keywords
Digital curation, digital preservation, electronic record preservation, e-mail list archives, e-mail preservation, trustworthy digital repository

1. INTRODUCTION
H-Net—an international consortium of scholars and teachers with the oldest collection of born-digital and content-moderated arts, humanities, and social science material on the Internet—is hosted by MATRIX, a digital humanities research center at Michigan State University. MATRIX received a grant from the National Historical Publications and Records Commission (NHPRC) to assess and improve upon the preservation practices for the H-Net e-mail lists. This collection of more than one million e-mail messages on more than 180 public networks and more than 230 private (administrative) lists is considered a valuable scholarly resource. The Center for Research Libraries (CRL) and Online Computer Library Center (OCLC)’s Trustworthy Repository Audit & Certification (TRAC): Criteria and Checklist was the primary tool used in the assessment. As a testbed for use of the TRAC, this H-Net preservation project will interest archivists and other information professionals who manage large collections of electronic records.

2. HOW H-Net WORKS
H-Net runs on LISTSERV software. Users subscribed to a public e-mail list send messages in plain text, with no attachments, to the list editor for approval. Messages may take from a few seconds to several days to post after approval, at which point they are stored in files called “notebooks.” These notebooks are concatenations of messages posted during a seven-day time period. A log browse cache extracts key metadata and creates MD5 hashes for each message, and this metadata is written to a database cache. A web browser interface allows users to retrieve individual messages, each of which is identified by the system through a combination of a notebook file name and an individual message’s MD5 hash.

3. PRESERVATION ASSESSMENT
3.1 Use of the TRAC
The Trustworthy Repository Audit & Certification (TRAC): Criteria and Checklist was used to assess the preservation practices for the H-Net e-mail lists. This TRAC includes three sections: Organizational Infrastructure; Digital Object Management; and Technologies, Technical Infrastructure, and Security. Each section consists of a number of criteria that require supporting documentation as proof of fulfillment. Institutions wishing to become third-party repositories for other organizations’ digital archives may use the TRAC to establish the requisite credentials. In the case of H-Net, MATRIX used the TRAC for self-assessment to ensure its efficacy as a preservation environment and to highlight areas that required improvement.

3.2 Existing Preservation Practices
3.2.1 OAIS Model
The H-Net preservation system follows the Open Archival Information System (OAIS) model, a reference model for an archive that has accepted the responsibility to preserve information for a designated community.

3.2.2 Backup and Storage
At the time of assessment, the backup and storage system for H-Net and MATRIX as a whole was based on daily incremental and weekly full backups to tapes that were cycled through the system approximately every six weeks and replaced as needed. Monthly full, “permanent” backup tapes were also made and kept in a secured room.
4. PRESERVATION IMPROVEMENTS

4.1 Backup and Archival Storage
Backup and storage improvements for MATRIX include establishing reciprocal and onsite storage arrangements; creating more than one set of “permanenr” backup tapes; putting all “permanent” backup tapes on a retention schedule; and establishing a backup log. Archival copies of the H-Net data will be made annually to tapes that will be refreshed every five years, and “dark” and distributed archival storage options are being explored.

4.2 Authenticity

4.2.1 Fixity
Fixity must be established for individual messages on submission and for notebook files on creation. If calculated at time of ingest, the messages may use the same MD5 hashes generated for discovery purposes for fixity checks. SHA-2 message digests will be calculated for the notebook files. Message hashes will be validated at the time of notebook completion, with notebook hashes validated on a weekly basis. Digital signatures will be generated for each list and updated and validated as needed to ensure against notebook deletion.

4.2.2 Other Authenticity Measures
Notebook modification rights will be restricted to authorized MATRIX personnel.

4.3 Attachments
Preservation of attachments to messages on the private lists must begin with providing browser access to those lists. Conversion tools will be kept in reserve or pointers to websites containing conversion tools provided for the most common formats. A technology watch will be established to keep up with the availability and usage of new formats and versions.

4.4 Other Improvements

4.4.1 Technical
Links within messages will be preserved and redirected to archived websites as needed. Shorter persistent URLs for messages will be mapped to the actual URLs, easing the citing of H-Net messages by researchers.

4.4.2 Administrative
A succession plan is being negotiated with another institution, in the event that there comes a time when MATRIX can no longer host H-Net. Supporting documentation for the criteria laid out by the TRAC checklist is being gathered and policies created to ensure the soundness of H-Net as a preservation system and trusted digital repository.

5. ACKNOWLEDGMENTS
MATRIX and Michigan State University thank the National Historical Publications and Records Commission (NHPRC) for funding this project.

6. REFERENCES
Federal Libraries Digital Preservation Census

Deborah Keller
U.S. Army Corps of Engineers, HECSA Library
Alexandria, VA 22315 USA
deborah.eb.keller@us.army.mil

Ann Harrison
Library of Congress, FLICC/FEDLINK,
Washington, DC, 20540 USA
anha@loc.gov

ABSTRACT
In this poster session, we share the planning and early data collection stages of a multi-year effort to survey the federal library community about its digitization efforts. We will discuss the overlapping sub-communities within the federal library sector which make it challenging to break this survey project into manageable pieces. We will discuss efforts to market this survey initiative so that we can encourage participation and show benefits to libraries for assisting with this information collection. We will also describe strategies that are developed for achieving maximum response rates and avoiding duplicate or unsanctioned responses. We will share survey questions that probe not only what digital projects are being undertaken in libraries, but the methods, funding, staffing, and work practices that are involved in these initiatives. While we intend to present some of the preliminary data collected from the first round of libraries surveyed, the final goal, an online directory of digital projects taking place in federal libraries, will be premature at the time of this presentation.

Keywords
Federal libraries, digitization, preservation, project management, resource management, knowledge management

1. INTRODUCTION
The Federal Library and Information Center Committee (FLICC) Preservation and Digitization Working Group is an ad hoc working group of library professionals that has been organized to develop strategies for the long-term preservation and access to federal library resources. This committee has a multi-faceted mission. It will work to assess the preservation and digitization needs of federal libraries and their users, develop recommendations and policies for digitization programs to ensure the responsible stewardship of federal digital assets as they relate to persistence, accessibility, and usability, and to identify and recommend standards and best practices for digitization projects, policies, and programs. The working group will also champion the needs for preservation and digitization throughout the Federal community, to advocate and educate on behalf of library disaster planning, and to foster and facilitate outreach and partnerships between federal libraries and with other preservation-minded organizations.

In order to accomplish a mission this far-reaching, the FLICC Preservation and Digitization Working Group has established subcommittees. One of these is focusing specifically on the information-gathering efforts of the working group. The group has adopted the mission of assessing the federal library community to determine what digitization projects are currently being undertaken, the infrastructure that is being used to support these efforts, and the ongoing needs of the libraries as they pursue digital initiatives. The findings of this information gathering will be consolidated and used to provide access to digital materials.

There have been several efforts to collect data about digitization projects underway within federal libraries in the past few years. None of these has been completely successful and many federal librarians still feel that they do not know what digital projects are being undertaken at other libraries. In some cases, librarians are unaware of digital projects going on in other libraries within the same agency. This data collection effort intends to distinguish itself from previous survey efforts. First, it intends to collect data in an iterative fashion, continuing to update the information that is available to federal libraries and their users. Second, it intends to use active information gathering methods to ensure that there are not gaps in the data collected rather than relying entirely on voluntary responses.

2. METHOD
The committee will put together a questionnaire about federal libraries’ digitization efforts, using it as the basis of a two-pronged survey effort. The questions will be sent directly to a single, authoritative contact on a predefined list of federal libraries. Phone calls to these representatives will follow, ensuring that the survey team receives responses from all libraries on the list.

Because the federal library community is both large and diverse, the survey team believes that it is impossible to achieve a high response rate—near 100% of libraries—if all of them are surveyed simultaneously. Instead, the team has decided to survey beginning with a small, well-defined group of libraries, broadening this list with each iteration of data collection. FLICC represents federal libraries at all levels and as such, its members are a cross section of the federal library community. Prior to being sent to FLICC member libraries, the survey questions will be pre-tested by the libraries represented on the survey team.

What materials are being digitized is central to the purpose of the information collection efforts. However, the survey team is also interested in the infrastructure behind these digital projects and the methods being employed in them. The team intends to collect data on the money, human resources, and time devoted to the projects, digitization methods and software being used, standards...
to which the project adheres, and workflows that have been adopted. The team also hopes to assess qualitative aspects about the projects. For example, librarians will be asked to assess their agency’s commitment to digitization. They will be asked to describe any partnerships involving the digital projects and to enumerate key challenges that their projects have faced. They will also be encouraged to suggest how FLICC can better support the digitization efforts going on within federal libraries.

3. DISCUSSION

Surveying the entire federal library community about their involvement in digitization efforts is a tremendous undertaking. The survey subcommittee feels that the only way to measure progress in a project of this size is to break it into manageable pieces. The first of those is described above. At the same time, the team is working to place this phase within the context of the larger survey effort and to begin to map out a way ahead.

In a large, multi-year effort such as this one, it is essential to raise awareness of the effort beyond the survey team itself. To do this, the survey team will enlist the assistance of FLICC, taking advantage of its quarterly membership meetings to describe the survey efforts, raise awareness among libraries who have not yet submitted information about their digitization projects, and solicit the involvement of and feedback from the entire federal library community to ensure that continuing survey efforts collect information that is of value to the entire community, not just to the survey team. The next meeting in early December is ideally timed to alert FLICC members about the first round of surveying which will take place early in the new calendar year.

At future meetings, the team will update the community on the progress of the ongoing surveying and keep them informed about how to access the projects about which information has been collected. 2009’s first FLICC meeting will take place shortly after the first batch of data has been collected. It is anticipated that preliminary trends may already be apparent as data begins to be analyzed. The process used during the first round of data collection will also have been examined by the team and their conclusions used to select recipients for the next round of questionnaires.

The survey team also plans to piggyback its survey efforts on the work of other FLICC committees. For example, a recent subcommittee studied the anticipated physical space needs of federal libraries during the next ten years. The digitization team is studying the survey methods employed by that team, hoping to improve on them in order to get higher return rate of the information requested. At the same time, the team will also study their results, looking to libraries who anticipate losing space as early targets for the digitization survey. The survey team also anticipates feeding its early findings to other FLICC committees in order to better inform their planning. At this time, it is felt that the FLICC Education working group may be a valuable partner of the Preservation and Digitization working group, serving to fill in digitization training gaps identified during the survey process.

4. ACKNOWLEDGMENTS

Our thanks to the Federal Library and Information Center Committee (FLICC) for facilitating cooperation and information sharing between federal libraries and its support for outreach beyond the federal library community.
Extending an LIS Data Curation Curriculum to Include Humanities Data

Allen H. Renear, Lauren C. Teffeau, Patricia Hswe, Molly Dolan, Carole L. Palmer, Melissa H. Cragin, and John M. Unsworth
Center for Informatics Research in Science and Scholarship
Graduate School of Library and Information Science
University of Illinois at Urbana-Champaign
Champaign, IL 61820 USA
{renear, teffeau, phswe, madolan, clpalmer, cragin, unsworth}@illinois.edu

ABSTRACT
We describe an IMLS-funded project to extend an existing data curation curriculum to include humanities data.

Categories and Subject Descriptors
K.3.2 [Computing Milieux]: Computer and Information Science Education – curriculum, information systems education

Keywords
Data curation, humanities, education, curriculum, libraries

1. OVERVIEW
As the creation, storage, and manipulation of large stores of data becomes widespread among scholarly communities—playing a central role in contemporary research practices—the need for state-of-the-art data curation education becomes critical.

Although the data curation community has emphasized scientific data needs, disciplines in the humanities are also seeking ways to not only develop and preserve digital content but also to enhance and document this content in order to ensure interoperability and functionality across the disciplines, and over time. In particular libraries, museums, archives, and other institutions managing cultural content require new kinds of expertise.

An IMLS-funded project, at the Graduate School of Library and Information Science (GSLIS), at the University of Illinois, directly addresses this need. This project extends the Data Curation Education Program (DCEP), a concentration within the GSLIS ALA-accredited masters program, to include humanities data, training a new generation of LIS professionals qualified as humanities data curators and providing continuing education for practitioners currently in the field. The new expanded curriculum is designed to be a model for state-of-the-art data curation.

2. NEED
Humanities disciplines such as history, literature, art, music, and philosophy are profoundly important to our society. These disciplines simultaneously create and mobilize the documentary record of the human condition, helping us remember and understand who we are and what we may become. They are part of every level of education, and they continue, throughout our lifetimes, to enrich experience and inform decision-making. Moreover, these are also disciplines whose substance is at the heart of the now economically-central culture and entertainment industries. For centuries, libraries and museums have been in the business of preserving and providing useful access to this record.

The humanities are an information-based domain in a double sense: both the development of data on the one hand, and its exploitation in new scholarship, education, culture on the other, are intrinsic to these disciplines, and thoroughly intertwined:

Humanists have traditionally viewed locating and compiling information, or ‘data,’ as a basic task of scholarship [7].

From the start of the data lifecycle, curation measures must be in place not only to establish authority and prevent data corruption, deterioration, and loss, but to ensure maximum value in digital contexts [5, 8]. Data curation has been recognized as an e-science problem, because of the rapidly growing amount of computationally created and processed science data [4]. However, data curation is no less a problem for the humanities [3, 6]. Increasingly, humanities scholarship utilizes computing applications for the display, analysis, and understanding of, for example, works of art and literature, events in history, archeological discoveries, or linguistic text corpora—all of which have in turn stimulated the gathering of digital data:

[R]esearch in the humanities is becoming data-centric, with a large amount of data available in digital formats … [this] quickly change[s] the landscape of humanities research [1].

The new digital context — data, tools, practices — challenges the methods and techniques with which we pursue traditional curatorial objectives such as authenticity, provenance, authoritative reference, and annotation. Moreover it also challenges our understanding of those objectives. The need for improved understanding of curatorial practices, and education of a humanities data curation workforce is urgent.

This work is licensed under the Creative Commons Attribution-Noncommercial-No Derivative Works 3.0 Unported license. You are free to share this work (copy, distribute and transmit) under the following conditions: attribution, non-commercial, and no derivative works. To view a copy of this license, visit http://creativecommons.org/licenses/by-nc-nd/3.0/.

DigCCurr2009, April 1-3, 2009, Chapel Hill, NC, USA
3. THE PROGRAM
The specific additions to the DCEP curriculum will reflect the results of an ongoing needs analysis and consultation with our cooperating institutions. However the general outline is clear.

3.1 Approach
Four core themes structure our focus:
(i) text and documents
(ii) digital artifacts and media
(iii) collections, metadata, and ontologies
(iv) scholarly information use and behavior

Treatment of these topics draws directly on ongoing research of GSLIS faculty in metadata, information modeling, document processing, ontology development, archival management, and information storage and retrieval, as well as traditional LIS areas such as indexing and abstracting, classification, collections development, user studies, and service design and management.

Three general principles provide a distinctive orientation.

- The humanities disciplines have already evolved sophisticated curation practices and theories, and in the last 50 years these have been refined and extended to support computation and digital storage. This rich tradition must inform any further development of a humanities data curation curriculum.
- The curation of cultural information has much in common with scientific data curation, but distinctive features as well. The interaction between the two areas is reciprocal, each informing the other and together advancing the development of data curation as a discipline.
- Many professions and disciplines are contributing to the development of data curation, however library and information science is well-suited to provide an over-arching curricular framework [9].

3.2 Outcomes
- Documentation of best practices for humanities data curation.
- New courses and educational materials that address digital humanities needs, through GSLIS’s ALA-accredited masters of science program (http://www.lis.uiuc.edu/programs/ms/).
- A network of cooperating institutions for internship, practicum, and job placement for enrolled students.
- A common framework for representing data curation concepts and practices that applies both to scientific and cultural data.

All materials, including case studies and curricular materials will be made freely available via UIUC’s institutional repository.

3.3 Cooperating Institutions and Advisory Committee
Advisory board members will help identify information problems, best practices, case study topics, as well as facilitate field placements and review the completed curriculum. Advisory board members represent the following cooperating institutions:

- OCLC Online Computer Library Center. Lorcan Dempsey, Vice President and Chief Strategist
- The Perseus Digital Library, Tufts University. Gregory Crane, Editor-in-Chief
- The Brown University Women Writers Project. Julia Flanders, Director
- Centre for Computing in the Humanities, King’s College London. Harold Short, Director
- The Institute for Advanced Technology in the Humanities, University of Virginia. Daniel Pitti, Associate Director
- Unit for Digital Documentation, University of Oslo. Christian-Emil Ore, Director of Research

3.4 Future Events
- Summer Institute for Humanities Data Curation, May 18-22 2009: a weeklong workshop for academic and research librarians hosted by DCEP (http://cirss.lis.uiuc.edu/ColMeta/decp/SummerInstituteHumanities.htm).
- The iSchools Conference, February 3-6, 2010: hosted by GSLIS (http://www.ischools.org/).
- International Digital Curation Conference, 2010: hosted by GSLIS and the Digital Curation Centre, UK (http://www.dcc.ac.uk/).

4. ACKNOWLEDGMENTS
This program is sponsored by the Institute of Museum and Library Services (RE-05-08-0062-08).

5. REFERENCES

Data Management and Curation of Research Data in Academic Research Environments

Barrie E. Hayes  
Health Sciences Library  
Univ. of North Carolina at Chapel Hill  
Chapel Hill, NC, 27599 USA  
barrie_hayes@unc.edu

James L. Harroun  
School of Info. & Library Science  
Univ. of North Carolina at Chapel Hill  
Chapel Hill, NC, 27599 USA  
james.h@unc.edu

Brenda Temple  
Structural Bioinformatics Core Facility  
Univ. of North Carolina at Chapel Hill  
Chapel Hill, NC, 27599 USA  
brenda_temple@med.unc.edu

ABSTRACT
The Structural Bioinformatics Core Facility at the University of North Carolina at Chapel Hill (SBI Core) assists researchers university-wide in computational structural biology techniques and incorporating structural biology/bioinformatics into their grants and publications. The SBI Core works with a diverse population of researchers from numerous departments and provides support to an ever-changing body of research. The computational biology services provided by the SBI Core are data-intensive and use a diverse and distributed set of applications for processing, data storage, and data management.

As the amount of data and number of projects have increased, the SBI Core requires an effective strategy for managing data and facilitating data sharing between the SBI Core and the researchers it assists. The UNC-CH Health Sciences Library (HSL) has begun a collaborative project with the SBI Core to identify the crucial data management needs and to envision new roles for the library in e-science and data management. In partnership, the SBI Core and the HSL have identified major obstacles in data sharing, data management, and data access. Furthermore, the SBI Core and the HSL will develop solutions in which the library facilitates collaboration among campus resources and matches unmet needs to external resources. One of the library’s goals in this proof-of-concept project with the SBI Core is to become a central campus resource for research support and data management.

Categories and Subject Descriptors
K.4.3 [Organizational Impacts]: Computer-supported Collaborative work; J.3 [Life and Medical Sciences]

Keywords
Data management, research data, cyberinfrastructure, bioinformatics, computational structural biology, information and library science.

1. INTRODUCTION
Scientific research data is increasingly digital and extensive in size. Digital experimental data is characterized by variations in configuration parameters, systems used, and output formats that do not exist with print or textual data stored in laboratory notebooks. Effective and efficient management of these growing, complex, and computationally-generated data sets are a particular challenge for researchers in academic scientific research environments. One issue facing researchers is that grant funding typically does not require or fund a data management plan. A second issue is that much of the research data is primarily generated, organized and stored by graduate students and post doctoral fellows who regularly join and leave the research team. For this project, the director of the SBI Core is collaborating with information and library science students and professionals to analyze the data environment and workflows and to determine viable solutions for improving data management processes in that environment.

2. BODY TEXT
SBI Core Facility at UNC: Computational Research and Data Generation
The Structural Bioinformatics Core Facility at the University of North Carolina - Chapel Hill (SBI Core) works with a variety of research laboratories to perform X-ray crystallographic structure determination, Nuclear Magnetic Resonance spectroscopy, molecular dynamics simulations, and bioinformatic analyses for UNC research projects. Data from these various analyses is generated and managed primarily by graduate students and post doctoral fellows working for the particular research lab. There are frequently no mandated formats, processes or policies for access, storage, and retention of this data.

Challenges Curating, Sharing, and Archiving Computational Biomedical Data in an Academic Setting
Issues that commonly arise in this environment when graduate students and post doctoral fellows leave the institution are that the data the departing student/post doc was managing are often inaccessible, incomplete, or missing altogether, resulting in losses in valuable researcher time and research outcomes.

An additional factor is that although the SBI Core generates and stores the data, it is not directly affiliated with the specific research departments with which it works. The data is temporarily stored in this environment and needs to remain accessible to the principal investigator and lab team conducting
the research. Additionally, there is no adequate and secure institutional research computing storage space available for shared use by the lab team or by the SBI Core.

Cooperative Efforts Between the SBI Core and the UNC Health Sciences Library to Model Research Data Management Practices

Evidence suggests that these challenges are common in scientific research environments [1, 3, 6, 7]. In this case, the SBI Core Facility has assistance from the UNC Libraries in addressing these issues. To date, staff from the UNC Health Sciences Library has met with the SBI Core director to define and describe the facility context, specific data workflows, and the issues of concern. This team is working with a committee of campus science librarians to identify partners among campus Information Technology Services, the School of Information and Library Science, and other key information management experts. This project team will review data management heuristics, guidelines, best practices, institutional and external resources and policies.

The team is led by members of the Health Sciences Library and the SBI Core as an initiative to identify the opportunities created by e-science [2, 4, 5, 6]. Based on its findings this group will recommend software, process, and policy solutions to assist this Core Facility environment in standardizing and stabilizing data management practices. This prototype project with the SBI Core will function as proof-of-concept to extend services campus-wide and the team will use this model to work with other units on campus.

3. REFERENCES


Contextual Information from Blogs in Video Digital Curation

Robert G. Capra  
School of Information & Library Science  
University of North Carolina  
Chapel Hill, NC 27599 USA  
rcapra3@unc.edu

Rachael Clemens  
School of Information & Library Science  
University of North Carolina  
Chapel Hill, NC 27599 USA  
rclemens@unc.edu

Christopher A. Lee  
School of Information & Library Science  
University of North Carolina  
Chapel Hill, NC 27599 USA  
callee@email.unc.edu

Laura Sheble  
School of Information & Library Science  
University of North Carolina  
Chapel Hill, NC 27599 USA  
sheble@email.unc.edu

ABSTRACT
In this study, we examined the extent to which blog postings that provide links to digital videos also provide contextual information about those videos. Our analysis of blog entries that linked to YouTube videos about the U.S. presidential election revealed that most of the blog entries do discuss the videos to which they link but also contain content that is not directly related to the videos. Most of the blog entries provide some additional contextual information. This suggests that crawling and capturing blog pages that link to YouTube videos can serve as a means to gather contextual information about those videos. We suggest future research that will (1) focus on the types of contextual information provided by blog entries; (2) determine whether there is a small subset of blog entries that provide the majority of contextual information about videos to which they link; and (3) identify aspects of blog entries that have the potential to positively impact more efficient collection of the entries. Answers to these questions will help curators of digital collections to better identify and collect content from the blogosphere to provide contextual information that will help future users to make sense of videos in their collections, while minimizing the resources required to capture additional content from the Web.

Keywords
Digital curation, digital preservation, contextual information, blogs, blogosphere, videos, video collections.

1. INTRODUCTION AND PURPOSE
When building long-term digital collections, it is essential not only to ensure continuing access to “target digital objects” but also to create, capture and manage contextual information to allow future users to understand, make sense of, analyze and use the target digital objects [1]. The VidArch project is capturing YouTube videos and web pages associated with the election, as well as exploring strategies and building tools for curators of digital collections to appraise and describe such materials.

In the study reported in this poster, we examined responses to online videos related to the U.S. presidential election using two types of data: 1) videos posted on YouTube and 2) blogs indexed in Google Blogsearch and Technorati. For each video in our study, at least one blogger had sufficient interest in the video to provide a link to it from his/her blog. The nature of interest in the video could be expressed as humor, disdain, enthusiasm, or not explicitly expressed at all. Bloggers often provided background information and, in some cases, provided coverage and analysis of the video and its place in contemporary society that one might expect to find in a newspaper article about the video.

2. METHODOLOGY
We used ContextMiner (http://contextminer.com) to generate a corpus of videos related to the 2008 U.S. presidential election. From this collection we analyzed videos retrieved from search queries on Barack Obama, Mitt Romney, and Tom Tancredo (182 videos total). A previous study [1] provided a collection of blog pages that were retrieved through a systematic crawl of Google Blogsearch and Technorati results using similar search terms for the 2008 U.S. presidential election. From this set of blogs we captured a subset that provided a direct link to at least one of the videos in the three video sets (422 blog posts total).

Candidates were chosen based on their standings in presidential primaries at the time of data collection (January 2008). We chose the two leading contenders (Obama and Romney) and two underdogs (Tancredo and Vilsack). We do not include Vilsack in our findings, because of an insufficient number of blog entries with links to the Vilsack videos in our data.

<table>
<thead>
<tr>
<th>Question</th>
<th>Object of Analysis</th>
<th>Coding Categories</th>
</tr>
</thead>
<tbody>
<tr>
<td>Is the video about the candidate?</td>
<td>YouTube video</td>
<td>3 = about the candidate 2 = about the election but not the candidate 1 = about neither the candidate nor the election</td>
</tr>
<tr>
<td>What portion of the blog entry is about the video?</td>
<td>Blog entry that links to the YouTube video</td>
<td>3 = entire entry 2 = part of the entry 1 = not part of the entry (e.g. in sidebar) C = only in comments</td>
</tr>
<tr>
<td>To what extent does the blog entry provide contextual information related to the video?</td>
<td>Blog entry that links to a YouTube video</td>
<td>3 = provides substantial amount of contextual information (like a news item about the video) 2 = some contextual information beyond that provided by title of video itself 1 = no real contextual information</td>
</tr>
</tbody>
</table>

Table 1. Coding Schema

This work is licensed under the Creative Commons Attribution-Noncommercial-No Derivative Works 3.0 Unported license. You are free to share this work (copy, distribute and transmit) under the following conditions: attribution, non-commercial, and no derivative works. To view a copy of this license, visit http://creativecommons.org/licenses/by-nc-nd/3.0/.

DigCCurr2009, April 1-3, 2009, Chapel Hill, NC, USA
We examined and coded a statistical sample of the blog pages for the three candidates. The authors engaged in four preliminary coding rounds to establish coding categories and help with inter-coder reliability. Table 1 shows the final coding schema used to rate the video and blog collections. The coding schema is based on three questions: First, is the video directly relevant to the initial candidate, the election more broadly, or neither? Second, what portion of the content of a given blog entry is directly related to the video to which it links? Finally, to what extent does the blog entry provide contextual information related to the video? Blog entries that linked only to videos that were coded as “about neither the candidate nor the election” were excluded from the analysis (i.e. they were not coded for quantity or extent) because the videos did not meet selection criteria for a collection about the 2008 U.S. presidential election. Each video and blog entry was coded by two different coders.

3. FINDINGS
Inter-coder reliability ratings indicated substantial agreement (Cohen’s kappa averaged 0.76 for quantity and 0.67 for extent). In cases where the two coders disagreed on a rating, we used the lower of the ratings in the findings presented here.

3.1 Video Relevance
Consistent with an earlier study [1], most of the videos returned from the Obama, Romney, and Tancredo search queries were relevant to the queries issued. Of the 182 videos analyzed, 75% were directly relevant to the candidate named in the query, 9% were directly relevant to the candidate named in the query, 9% were relevant to the election in general and 16% were no longer available (removed from YouTube for a variety of reasons).

3.2 Extent of Blog Post
As pictured in Figure 1, overall, 15% of the blog entries were entirely about the linked YouTube video, 52% were determined to be only partially about the linked YouTube video, and only 3% contained content linking to the YouTube video only in a sidebar on the blog page. Twenty-nine percent (29%) of the blog entries were no longer available on the Web.

3.3 Contextual Information about the Videos
One of the primary goals of our study was to determine whether or not blogs provide valuable contextual information. Overall, 47% of the blog entries were determined to provide some contextual information beyond simply the title of the video. Extensive coverage was less frequent. Only 7% of the blog entries provide in-depth or news-like coverage or analysis of a video; and 17% provided only a few words of contextual information.

4. IMPLICATIONS AND FUTURE DIRECTIONS
The harvesting of contextual information from external sources will become increasingly important for curators of digital collections as (1) items in environments such as YouTube play a significant role either in phenomena that should be documented or as documentary evidence of those phenomena; and (2) the environments themselves provide limited contextual information.

The VidArch project has also been running crawls on many other topics (e.g. energy, epidemics, health, natural disasters, truth commissions). We do not yet know the extent to which the findings in this paper will be relevant to efforts to document non-election phenomena. We also do not yet have generalizable findings about the likely rate of diminishing returns when collecting additional contextual information.

In future research, we will focus on the specific types of contextual information provided by blogs (e.g. information about the event documented in a video, identity of its creator, relationships with other videos). We will also attempt to further characterize the blog sources that we have found to consistently yield a substantial amount of contextual information. For example, do the blog pages that are most rich in contextual information tend to come from specific domains, have high in-link counts, have longer average posting length, or simply come later chronologically (thus having the benefit of building on facts revealed by other blogs)? Answers to these questions will help us to formulate rules and heuristics for collecting blogosphere content that will have a high probability of helping future users to make sense of the videos in a collection, while minimizing the resources required to capture additional content from the Web.

5. REFERENCES
What Should We Teach about METS in a Digital Preservation Course?

James Waters
College of Information Science and Technology
Drexel University
Philadelphia, PA 19104 USA
james.waters@drexel.edu

Robert B. Allen
College of Information Science and Technology
Drexel University
Philadelphia, PA 19104 USA
rba@drexel.edu

ABSTRACT
We describe a METS (Metadata Encoding and Transmission Standard) teaching assignment which forms part of a foundational Digital Preservation course for an MS(LIS) degree. The assignment requires students to both critically evaluate this important framework and apply it practically to metadata management for digital objects. The results indicate that it was a valuable assignment for LIS and IS students who could conceptually grasp METS readily but that many have trouble with integrating external metadata schemes and with XML syntax. These results are informing a redesign of the assignment.

Categories and Subject Descriptors
K.3.1 [Computers and Education]: General

Keywords
Education, Metadata, METS, OAIS, Preservation.

1. OVERVIEW
The Drexel Master's-level Digital Preservation course attempts to provide a comprehensive overview of digital preservation for both Librarians and IS professionals. This course teaches putative library and IS professionals theoretical and pragmatic issues involved in the selection, appraisal, ingestion, curation and management of digital collections. This course bridges the boundaries between conventional archival and library sciences and the emerging issues of digital preservation. The majority of students are enrolled in the College's MS(LIS) program. This was an advanced course and the students already had some exposure to XML, metadata, and repository management.

The course covers the underlying preservation principles common to both physical and digital materials but expands on the different strategies, approaches and technologies required for dealing with digital materials. This course covers a broad range of repository issues ranging from toolkits (e.g., Archivist Toolkit, DSpace), to multimedia (e.g., MPEG-7), to service specification (e.g., LOCKSS and iRODS), but also for metadata which is applicable to digital repositories. In particular, we introduce METS (Metadata Encoding and Transmission Standard). METS is important as part of the OAIS framework both of which are growing in popularity. METS was an outcome of the Making of America 2 initiative [2].

The Metadata Encoding and Transmission Standard (METS) occupies a middle ground between the abstract package definitions derived from the OAIS framework and the low level data definitions found in the NISO extension scheme as illustrated in Figure 1. It provides a critical nexus between a conceptual framework and implementation issues. As the schema is implemented in XML, it also provides practice for students with high level technical validation issues.

Since we aim to help educate both future researchers and practitioners in this realm we are concerned that they should be equipped with the willingness to challenge accepted practices. Thus, we believe that students should be familiar not just with METS and OAIS on a conceptual basis but that they also have experience with the METS schema and practice with creating realistic METS documents in XML, a valuable skill.

By using a METS framework for a realistic task we hoped students would gain perspective as to what is and is not useful about METS. More importantly though we hoped to encourage them to get into the habit of challenging current wisdom early in their careers. METS imposes a structural framework for metadata creation, we wanted students to reflect on how useful was this structure in particular and structured approaches in general?

This work is licensed under the Creative Commons Attribution-Noncommercial-No Derivative Works 3.0 Unported license. You are free to share this work (copy, distribute and transmit) under the following conditions: attribution, non-commercial, and no derivative works. To view a copy of this license, visit http://creativecommons.org/licenses/by-nc-nd/3.0/.

DigCCurr2009, April 1-3, 2009, Chapel Hill, NC, USA

199
2. **THE ASSIGNMENT**

Students were asked to read some core readings including references [1] and [2] to get an understanding of the context for METS. Students are also given METS schema documentation and examples of METS XML documents for reference. Then, a short prologue describes the seven METS sections conceptually and continues:

```
For this assignment we will concentrate on:
Administrative Metadata Section, File Group Section, and Structural Map Sections,

Create a simplified METS document containing only the three sections mentioned above (Administrative, File, Structure Map) to describe a specific (named) Digital object. This object can be derived from a Book, Historical Document, Movie, Video, Painting, Photograph, Sculpture, other work of art, or Person.

Don’t get too worried about the precise XML syntax for this exercise but make sure the document describes the object. Since many METS elements/attributes are optional use your judgment and only include those that would be useful for your object

Discuss the METS scheme with respect to:
- How well does the METS scheme allow you to describe your target object?
- How useful would your document be for retrieval or playback applications with respect to your chosen object?
- Are there any elements/attributes you would add to the METS scheme to make it more useful?
- Are there any elements/attributes you would remove from the METS scheme?

Finally choose a specific example of a second (different) object type from the list above and briefly (1 to 2 paragraphs) discuss how well the METS scheme deals with that type of object.
```

Fig 2. Text of the assignment.

We decided to focus this assignment on three sections. (Administrative Metadata, File Group and Structural Map) The Descriptive Metadata section is frequently used in conjunction with external schemas such as MARC and the Behavior section is seldom used in practice and refers to interface issues beyond the scope of this course, thus these two sections were omitted from the assignment though they are covered in the course.

Similarly, we chose to concentrate on the overarching conceptual structure. The schema for a METS document is comparatively complex and while our students are familiar with XML they are not all equally technically competent and we have found that those with greater technical ability will chose to be more technically accurate anyway and those with weaker technical skills should not be distracted by syntax.

3. **RESULTS**

Nine students completed the assignment. Six of the nine students understood the ability of METS to use wrappers for other metadata schemata to overcome limitations in the base scheme. One however did not define the external metadata scheme. While most students understood the relationship between the `structmap` and the `file section` on a conceptual level there were notable weaknesses found here. Some students created file pointers that did not agree with files in the file section. Some students incorrectly nested Divs, others created maps which simply listed a structure without relating it back to the file section. Many students had trouble understanding the use of the area element including confusion over how to label, begin and end areas, sometimes having literals for time segments. Other students created area elements representing the entire object or had beginning and end points that were inappropriate labels such as literal labels for parts of an image not coordinate pairs.

Several students remarked that the assignment was difficult and the XML syntax tricky. One student felt that METS offered no benefit above “Title, Subject, Author” for retrieval, most others felt the METS schema was intrinsically flexible and powerful due to the ability to pull in external schemata as required. Some though failed to understand that many METS elements were optional; despite being informed so explicitly. Similarly METS’ strength for representing complex or multi-representational objects was noted but sometimes METS was criticized for being difficult to apply to simpler objects. Overall students tended to be moderately positive towards METS.

Overall we consider the exercise to have been a moderate success. While students had problems implementing the METS schema they generally grasp how it worked conceptually. It seems that METS which was part of one week’s teaching required rather more coverage and this will be reflected in future use of this material.

4. **IMPLICATIONS AND EXTENSIONS**

This assignment represents one facet of our broader effort to introduce archival repositories and their metadata issues. As implied by Fig 1, METS is based on OAIS. Future plans include the use of JHOVE [4] for validating METS documents, specifically METS-ALTO documents and an exercise relating to Z39.87. The Z39.87 exercise will require students to both create metadata records for digital still images and critically evaluate the usefulness of the framework with respect to organization, and retrieval issues.

5. **ACKNOWLEDGEMENTS**

This project was supported by the Institute of Museum and Library Services (IMLS) Grant RE-05-05-0085-05 on developing a Model Curriculum for the Management of Digital Information.
6. REFERENCES


Sustaining Digital Preservation Organizations: What Discourse Analysis Can Tell Us about Market Demand and Long-Term Survival

Lori Eakin
School of Information and Library Science
University of North Carolina at Chapel Hill
Chapel Hill, NC 27599 USA
lorraine.eakin@unc.edu

ABSTRACT
In this poster, I show the results of a study in which the coevolution of community discourse and organizational structures in preservation and archiving have influenced the current ecology of the field. In particular, I use latent semantic analysis (LSA), a computational technique originating in the Information Retrieval field, to show how the historical trends, technological developments, values and professional issues within a community lead to a demand for organizational attributes that are eventually instantiated by the emergence of new forms and sub-forms of organization. The results of the analysis can then provide the historical and evolutionary basis for explaining why particular types of digital preservation forms have emerged and what types of hypotheses about the future sustainability of these forms can be derived.

Categories and Subject Descriptors

Keywords
Sustainability, organizational theory, digital archives, digital preservation, coevolution, community ecology, latent semantic analysis.

1. INTRODUCTION
We all rely upon the authenticity of documents such as birth certificates, minutes of political meetings, census documents, and records of taxation. Until the “digital age,” archivists had a relatively well-formulated process to provide assurances that as such documents and artifacts changed hands they remained true to their original in both content and form. Since the mid-1990s, however, the entire field has faced a series of environmental shocks. When documents are “born digital,” (i.e., created online), the ability to manipulate them, either deliberately or accidentally, rockets. Rapid obsolescence of technology means that the “original” as created can never be saved over time; files and file systems must be migrated to new technical environments on a regular basis, increasing the risk of degradation (or deliberate corruption). As more and more people have become accustomed to accessing cultural materials on the World Wide Web, the public has come to expect that such objects be not only well preserved, but also rapidly available upon demand. Government regulations regarding information availability during litigation have led to the need to save more information than ever before and to be able to retrieve relevant information quickly. Finally, on top of these changes, the archival mandate to save this information in perpetuity is held by organizations that are typically funded via soft money for periods of only two to five years.

These technical, social, legal, and economic factors have placed incredible stresses upon the field, leading to a number of exploratory attempts to create organizational structures that can handle all of the new requirements without losing focus on the original archival goals of information authenticity and integrity in perpetuity. A series of new organizational forms and sub-forms has arisen, from digital libraries and digital archives to institutional repositories, web archiving organizations like the Internet Archive, and personal information archiving organizations like Wuala.

A key question that arises from all of this is, how sustainable are these forms of archiving?

Sociologist Martin Ruef has argued that one can use discourse analysis to begin to understand the capacity of a community for accepting new forms of organization (2000). He provided such an explanation for the health care field by showing how the demands for attributes related to potential organizational forms led to the later development of actual forms. To do this he used latent semantic analysis (LSA) to assess the evolving issues, values, and trends within the health care field over time. He combined this discourse analysis with methods from community ecology that suggest new forms of organization will develop as a result of community demands until the “carrying capacity” of the form’s niche has been met. (The carrying capacity is the maximum number of organizations that a particular environmental niche can support.) From this analysis Ruef was able to provide both a descriptively rich explanation of the coevolution of the health care community and the development of new organizational forms.
within that community and to offer testable hypotheses about the likely survivability of those organizational forms.

2. METHODOLOGY

LSA is an indexing technique that took root in the Information Retrieval world, where it has become a widely used technique for analyzing a corpus of documents to extract common conceptual strands and to cluster subsets of them around these concepts, even when the actual term representing the concept does not exist in a given document within the subset (Landauer et al. 1998). It has been used extensively not only in information retrieval problems, but also for language learning, psychological assessment of cognitive development, and essay test grading, where it has been shown to perform comparably with actual human graders (Landauer and Dumais 1997). It has also recently been used to examine the changing intellectual concerns of professionals within a given field to reflect the problems they face over time, their values, and their demands for solutions to ongoing issues in the field (Ruef 2000; Sidorova et al. 2008).

LSA operates by representing word meanings as vectors in k-dimensional space. In a technique related to factor analysis and multi-dimensional scaling, LSA fits terms together simultaneously into a space of k-dimensions (i.e., the semantic space). It does this by assuming that terms that tend to be used together in discourse reside relatively close together in such a semantic space. The closer the terms in this space, the more closely related (psychologically) they are in usage. It functions by breaking down every passage of text in a corpus into its component words and feeding those words into a matrix in which every word represents a row and every document represents a column. Each cell contains the number of times this word occurs in the passage. Singular value decomposition is then applied to the resulting matrix, solving the set of simultaneous equations that the matrix represents and generating a set of vectors in semantic space. The more specialized the corpus used, the greater the likelihood that a smaller number of dimensions will be sufficient for generating meaningful results (Ruef 2000). In fact, the “appropriate” number of dimensions to represent the semantics of the matrix adequately is empirically derived, and a researcher who wishes to use this technique to represent meaning generally must adjust the dimensionality in a series of “tests” to find the best fit for his or her research purposes.

My project uses latent semantic analysis to structure the discourse of information science professionals, in particular, archival and English speaking preservation professionals from 1985 to the present. Utilizing a corpus of published professional journals and conference proceedings available on the World Wide Web, a breakdown of the key concepts, issues, and values are identified, and an analysis of their evolution over the past twenty-three years occurs. The results of this discourse analysis are then mapped historically to the development of archival organizational forms.

3. RESULTS

As of this writing, the material used to generate the corpus has been collected and pre-processing is being initiated. When the pre-processing is finished, the analysis will take place, generating a set of concepts representing the key concerns of the archiving and preservation field over the past twenty-some years. This information will then be available for a number of types of analysis. It can, for instance, be compared historically to the primary organizational forms to assess the temporal patterns of organizational change over time. Also, with additional data on the formation and dissolution of organizations within each form, statistical analysis will enable one to estimate whether these forms have reached their carrying capacity and to test ecological predictions about their likely survivability and about potential evolutionary changes that could be expected within the field. (For example, once a form’s carrying capacity has been reached, one would normally expect that demands for new organizational features will be met, not by the entry of new organizations, but by the addition of these new features to currently existing organizations.)

4. REFERENCES


---

1 LSA is also able to fit terms to documents and documents to documents.
Getting the Tar Off Our Heels: 
Moving Forward with Archiving 
University of North Carolina at Chapel Hill Websites

Lisa Gregory
School of Information & Library Science
University of North Carolina at Chapel Hill
Chapel Hill, NC  27599 USA
gregoryl@email.unc.edu

ABSTRACT
During the 2008-2009 school year, the author, a DigCCurr fellow assigned to The University of North Carolina at Chapel Hill (UNC) University Archives and Records Management Services (UARMS), is exploring the feasibility of integrating website archiving into the UARMS workflow. Because this is a relatively new area for the UARMS, it will be vital to discover what is currently being done in the field to help UARMS situate its goals and resources into workflows that already have been developed and to learn from the difficulties that others have overcome. From there, test implementation of selected tools will be carried out to determine their feasibility for long-term use at UARMS. This poster will present the results of the author’s research and test implementation of different methods for archiving websites. As such, it will be of interest to those curious about how an institution can survey its resources and the field of available tools, and begin a web archiving program tailored to its needs. It will also showcase open source tools that the UNC UARMS explores/implements on a trial basis.

Categories and Subject Descriptors
H.3.5 [Information Storage and Retrieval]: Online Information Services—Web-based services; H.3.7 [Information Storage and Retrieval]: Digital Libraries—Collection

Keywords
Archiving websites, digital curation, digital preservation, university archives.

1. INTRODUCTION
Cultural heritage entities all over the world have watched the information they once collected in paper format accumulate online. Should format inhibit collection? Many institutions are increasingly answering that question in the negative. Integrating the archiving of websites into a long established collecting history is the mission of initiatives at the Bibliothèque nationale de France (BnF) [1], the National Library of Australia (NLA) [2-3], and the Library of Congress (LC) [4], to name a very few. The University of North Carolina at Chapel Hill (UNC) University Archives and Records Management Services (UARMS) is in a similar position, exploring the possibility of developing its own web archiving activities as a supplement to and continuation of its current collecting mandate. During the 2008-2009 school year, the author, a DigCCurr fellow assigned to UARMS, is exploring the feasibility of integrating website archiving into the UARMS workflow.

2. WEB ARCHIVING IN CONTEXT
Because websites are a relatively new format for the UARMS, it has been useful to survey the state of web archiving at archives and libraries. Staff from a number of institutions have published details about their efforts in beginning and sustaining web archiving activities. Many of these institutions are on the national level and take the breadth of their responsibilities to heart. Archiving on a national scale has been described within the broad directive of collecting a country’s “digital memory” [1] or “digital heritage” [5]. The Library of Congress believes that the “traditional functions of acquiring, cataloging, preserving and serving collection materials of historical importance to the Congress and the American people to foster education and scholarship extend to digital materials, including Web sites” [6]. The predilection for harvest on a national level seems to be mixed method, full-domain harvesting.

For state archives, the evidential value of websites and demands of public records laws drive website harvest and preservation. The proliferation of documentation on the web has even been called a “boon for the state library” [5]. Fewer reports are available of in-house efforts at universities, perhaps because their motivating user community is more targeted and their conceivable harvest scope smaller. Finally, for many disparate entities (and for a fee), the Internet Archive is providing this service through Archive-It (see [8] for a list of partners). It is clear that librarians, archivists, and records managers are pursuing a number of open source and fee-based solutions for a variety of web archiving initiatives. Their published accounts help UNC’s UARMS situate its goals and resources into workflows that have already been developed.

This work is licensed under the Creative Commons Attribution-NonCommercial-No Derivative Works 3.0 Unported license. You are free to share this work (copy, distribute and transmit) under the following conditions: attribution, non-commercial, and no derivative works. To view a copy of this license, visit http://creativecommons.org/licenses/by-nc-nd/3.0/.

DigCCurr2009, April 1-3, 2009, Chapel Hill, NC, USA
3. WEB ARCHIVING TOOLS
Although, as mentioned above, many institutions are taking advantage of the Internet Archive's web archiving infrastructure and expertise, for a number of institutions, this option is not feasible, especially on an experimental level. Instead, they are turning to increasingly sophisticated and user friendly open source tools. Published accounts of web archiving initiatives have detailed the rise and fall of open source and in house tools [9]. Currently, many have gravitated to the suite of options recommended by the International Internet Preservation Consortium [10]. Many of these tools have been used individually to support different steps, such as ingest or access, in the web archiving lifecycle. On the other hand, multi-function tools like the Web Curator Tool and NetArchive Suite offer users a packaged product that supports a number of different parts of the archiving workflow.

4. EXPERIMENTATION AT UARMS
The goals of this fellowship are to (1) identify web archiving tools by reviewing the literature, (2) test a number of these tools in order to determine their feasibility for UARMS, (3) record and propose possible web archiving workflows for UARMS, and (4) propose resource requirements should UARMS wish to archive websites on a permanent basis. A review of the literature discussing worldwide initiatives and widely used tools will inform UNC's UARMS about possibilities to explore, challenges to expect, and pitfalls to avoid. The author will be testing several of the tools mentioned above, targeting higher profile UNC websites whose paper documents are currently collected by UARMS. In addition to looking at the harvesting process, the author will also be experimenting with the Integrated Rule-Oriented Data System (iRODS™), a data grid software system, as a possible preservation storage environment. The author hopes that these efforts will result in website archiving workflow and resource suggestions for the UNC UARMS.

This poster will present the results of the author's research and test implementation of different methods for archiving websites. As such, it will be of interest to those curious about how an institution can survey its resources and the field of available tools, and begin a web archiving program tailored to its needs. It will also showcase open source tools that the UNC UARMS explores/implements on a trial basis.

5. ACKNOWLEDGMENTS
The author wishes to thank Drs. Helen Tibbo and Cal Lee, DigCCurr faculty advisors, as well as Janis Holder, UNC University Archivist.

6. REFERENCES
AUTHOR INDEX

Abbas, June ............................................. 158
Abrams, Stephen ....................................... 68
Agnew, Grace .......................................... 117
Albani, Sergio ........................................... 130
Allen, Robert B. ....................................... 199
Anderson, Sheila ...................................... 161
Angelis, Stavros ....................................... 170
Ashley, Kevin .......................................... 114
Ball, Alexander ........................................... 68
Beaudoin, Joan E. ...................................... 87
Becker, Christoph ....................................... 45
Beruti, Vincenzo ....................................... 130
Botticelli, Peter ........................................... 70
Bowden, Heather ....................................... 44
Bradley, Jana .......................................... 70
Bradley, Jim ............................................. 156
Bragg, Molly ............................................. 112
Bryner, Kyle Elizabeth .............................. 149
Burton, Adrian .......................................... 24
Callery, Bernadette G. .............................. 143
Capra, Robert G. ....................................... 196
Carrier, Sarah .......................................... 68
Chace, Myron B. ........................................... 60
Chapman, Stephen ...................................... 163
Chua, Mason ............................................. 57
Clemens, Rachael ....................................... 196
Conrad, Mark ............................................. 4
Constantopoulos, Panos ............................ 170
Conway, Paul ............................................. 7
Cox, Richard J. .......................................... 102
Crabtree, Jonathan .................................... 57, 90, 163
Cragin, Melissa H. ................................. 83, 92, 117, 191
Crawford, Claretta .................................... 85
Cruse, Patricia ........................................... 68
Cunningham, Richard ................................. 119
Davidson, Joy ........................................... 83
Day, Michael ........................................... 68
Dolan, Molly ............................................. 191
Dreyer, Malte ........................................... 68
Dryden, Jean ........................................... 19, 117
Duff, Wendy ............................................. 119
Dunn, Christina ........................................ 85
Eakin, Lori .............................................. 202
Eaton, Fynnnete ...................................... 163
Elford, Douglas ...................................... 73, 76
Eschenfelder, Kristin R. ........................... 117
Esteva, Maria .......................................... 87
Eubank, Kelly .......................................... 31
Fachry, Khairun Nisa ................................. 49
Fox, Edward A. ........................................ 2, 112
Fresko, Marc .......................................... 37
Frick, Caroline ......................................... 42
Fulton, Bruce .......................................... 70
Fusco, Luigi ............................................. 130
Gavrilis, Dimitris .................................... 170
Giaretta, David ...................................... 130, 177
Gisson, William Bradley ........................... 110
Gordon, Kelly M. ................................... 185
Grace, Stephen ....................................... 161
Granda, Peter .......................................... 1
Grasso, Chris .......................................... 1
Greenberg, Jane ....................................... 68
Gregory, Lisa ........................................... 204
Grindley, Neil .......................................... 6
Halbert, Martin ....................................... 90
Hancock, Beverlye H. ............................... 149
Hank, Carolyn .......................................... 2
Harrison, Ann ........................................... 189
<table>
<thead>
<tr>
<th>Name</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unsworth, John M.</td>
<td>191</td>
</tr>
<tr>
<td>Urbani, Carlo</td>
<td>154</td>
</tr>
<tr>
<td>Van de Sompel, Herbert</td>
<td>94</td>
</tr>
<tr>
<td>Van Diessen, Raymond J.</td>
<td>44, 114</td>
</tr>
<tr>
<td>Vidrine, M. Robert</td>
<td>149</td>
</tr>
<tr>
<td>Walters, Tyler</td>
<td>90</td>
</tr>
<tr>
<td>Ward, Jewel</td>
<td>57</td>
</tr>
<tr>
<td>Warwick, Claire</td>
<td>119</td>
</tr>
<tr>
<td>Waters, James</td>
<td>199</td>
</tr>
<tr>
<td>White, Kelvin L.</td>
<td>158</td>
</tr>
<tr>
<td>Whittington, Stephen L.</td>
<td>149</td>
</tr>
<tr>
<td>Wilczek, Eliot</td>
<td>4</td>
</tr>
<tr>
<td>Wildemuth, Barbara M.</td>
<td>2</td>
</tr>
<tr>
<td>Williams, Kathleen</td>
<td>6</td>
</tr>
<tr>
<td>Winget, Megan</td>
<td>42</td>
</tr>
<tr>
<td>Wojcik, Caryn</td>
<td>4</td>
</tr>
<tr>
<td>Wurl, Joel</td>
<td>6</td>
</tr>
<tr>
<td>Yakel, Elizabeth</td>
<td>7</td>
</tr>
<tr>
<td>Yang, Seungwon</td>
<td>2</td>
</tr>
<tr>
<td>Zhang, Junte</td>
<td>49</td>
</tr>
</tbody>
</table>