NEW SKILLS for a DIGITAL ERA

A Colloquium sponsored by
National Archives and Records Administration
Society of American Archivists
Arizona State Library, Archives and Public Records

31 May – 2 June 2006
Washington, DC

Proceedings edited by
Richard Pearce-Moses and Susan E. Davis

Society of American Archivists
17 North State Street, Suite 1425
Chicago IL 60602-3315 USA
312.606.722 Toll-Free 866.SAA.7858
Fax 312.606.0728
www.archivists.org

TABLE OF CONTENTS

Acknowledgements / v

Preface – Allen Weinstein / xii

Foreword / vii

Knowledge and Skills Inventory – Richard Pearce-Moses and Susan E. Davis / 1

Reflections – Richard Pearce-Moses and Susan E. Davis / 33

Appendix 1: Keynote address – Margaret Hedstrom / 35

Appendix 2: Case Studies / 41

Acquisition (Selection and Surveys, Transfer and Ingest)
- Eliot Wilczek and Kevin Glick, "Changes in Acquisition: A Guide to the Ingest of Electronic Records" / 41
- Ann Marie Przybyla and Geof Huth, "Conducting an Inventory of Electronic Records" / 47

Processing (Arrangement, Classification, Description)
- Timothy Pyatt, "Acquisitions: Assessment, Scheduling, and Transfer of Public Affairs Records" / 53
- Catherine Stollar and Thomas Kiehne, "Guarding the Guards: Archiving the Electronic Records of Hypertext Author Michael Joyce" / 57

Storage and the Digital Stacks
- Patricia Galloway, "The Eyes of Texas: What Can Archivists Learn from Working with a Digital Institutional Repository?" / 65
- Jennifer King, "George Washington University's Special Collection's Transformation into a Repository with Digital Services" / 73

Preservation
- Maria Esteva, "Text and Bitstreams: Appraisal and Preservation of a Natural Electronic Archive" / 77

Reference and Access
- Beth Yakel and Polly Reynolds, "The Next Generation Finding Aid: The Polar Bear Expedition Digital Collections: A Case Study in Reference and Access to Digital Materials" / 87
- Margaret Adams, "Archival Reference Services for Digital Records: Three and a Half Years Experience with the Access to Archival Databases (AAD) Resource" / 95

Managing Digital Archives: Balancing Responsibilities and Skills
- Rich Dymalski and Jerry Kirkpatrick, "One County's Attempt to Move from Zero to Digital in Record Time" / 105
- Cole Whiteman, "Mapping Processes in Motion: Practical lessons from the experience of discovering, visualizing, analyzing, and redesigning a complex process of digital archiving and dissemination" / 123

ACKNOWLEDGEMENTS

The authors wish to recognize the individuals who helped make the colloquium a success.

Advisory Board

Nancy Beaumont, Society of American Archivists
Lewis Bellardo, National Archives and Records Administration
Thomas Brown, Academy of Certified Archivists
Mary Chute, Institute of Museum and Library Services
Susan Davis, College of Information Studies, University of Maryland
Adam Jansen, Washington State Archives
Michael Kurtz, National Archives and Records Administration
Deanna Marcum, Library of Congress
Robert S. Martin, Texas Women's University
Susan McKinney, ARMA International and University of Minnesota
Karl Niederer, Council of State Archivists
Cheryl Pederson, ARMA International and Cargill
Richard Pearce-Moses, Society of American Archivists and Arizona State Library, Archives and Public Records
Roberta Schaffer, Library of Congress
Vernon Smith, National Archives and Records Administration
Robert Spindler, University Libraries, Arizona State University
Sharon Thibodeau, National Archives and Records Administration
Allen Weinstein, National Archives and Records Administration
GladysAnn Wells, Arizona State Library, Archives and Public Records

Planning Committee and Facilitators

Lucy Barber, National Archives and Records Administration
Susan Davis, College of Information Studies, University of Maryland
Katherine Deiss, Metropolitan Library System
Robert Horton, Minnesota Historical Society
Susan McKinney, ARMA International and University of Minnesota
David McMillan, National Archives and Records Administration
Richard Pearce-Moses, Society of American Archivists and Arizona State Library, Archives and Public Records
Robert Spindler, University Libraries, Arizona State University
Kathleen Williams, National Archives and Records Administration

Case Study Presenters

Margaret Adams, National Archives and Records Administration
Rich Dymalski, Maricopa County, Arizona
Maria Esteva, University of Texas at Austin

Patricia Galloway, University of Texas at Austin
Kevin Glick, Yale University
Geof Huth, New York State Archives
Thomas Kiehne, Fosforus; formerly Harry Ransom Center, University of Texas at Austin
Jennifer King, George Washington University
Jerry Kirkpatrick, Maricopa County, Arizona
Ann Marie Pryzbyla, New York State Archives
Timothy Pyatt, Duke University
Polly Reynolds, University of Michigan
Catherine Stollar, Harry Ransom Center, University of Texas at Austin
Cole Whiteman, ICPSR (Inter-university Consortium for Political and Social Research)
Eliot Wilczek, Tufts University
Elizabeth Yakel, University of Michigan

Recorders

Steven Bookman
Erin Greenwell
Brad Houston
Lisa Maxwell
Sara Muth
Linda Reib (née Colwell)

Special thanks to Allen Weinstein, Archivist of the United States, and to GladysAnn Wells, Director and State Librarian of the Arizona State Library, Archives, and Public Records, for their support for the colloquium. And to David McMillan of the National Archives and Records Administration, for his patience and perseverance in helping with logistics and planning.

FOREWORD

The work of archivists, librarians, and records managers is changing at a fundamental level. The very materials with which these professions work are undergoing a radical transformation. For decades, these information professionals[1] have handled a combination of physical documents composed of text formed with ink on paper, photographs made of silver or dye on papers and plastics, or other familiar formats. Archivists have also worked with machine-readable materials such as audio recordings (records, wire recordings, tapes, and compact discs) and moving images (film, videotape, and DVDs), but these formats were a small portion of the collections and often considered a specialty. Many archivists put these odd materials in a separate box at the end of the collection, treating them as tangential to the "real" records. The approach of concentrating on traditional formats and isolating the nontextual ones as supplementary is no longer valid.

The world has changed. Although no single event clearly dates the beginning of the digital era, several mark its rise. The development of ENIAC in the 1940s was a seminal moment.[2] The introduction of the personal computer in the 1980s made computing accessible to many people. And the web, made public in 1991, has made digital information seemingly ubiquitous.[3]

Archivists, in particular, have been concerned about digital information for some time. The Society of American Archivists established the Committee on Archival Records and Techniques in 1988 as a standing committee and a successor to the Automated Records and Techniques Task Force, an indication of the impact of this issue on the profession. For the past ten years, every president of the Society has mentioned the challenges of electronic records in his or her address to the members at the annual meeting. The National Archives and Records Administration began working with electronic records in 1968 and established the Center for Electronic Records in 1988. Still, many archivists ignored digital materials or saw it as an esoteric field of little relevance.

From 1989 to 1994 and again in 1996 and 1997, the National Association of Government Archives and Records Administrators sponsored the first organized effort aimed at helping government records professionals to develop programs aimed at managing electronic records.

---

[1] For sake of simplicity, the phrase "information professionals" (rather than the acronym "LARM," which was used in literature for the colloquium) will be used to describe librarians, archivists, and records managers collectively. Although many distinguish librarians as working with published materials, rather than records, those publications often function as records. Further, while the intent of this colloquium was to build consensus among these related information professions, most participants were archivists.

[2] See Scott McCartney, *ENIAC: The Triumphs and Tragedies of the World's First Computer* (New York: Berkeley Books, 1999).

[3] "Berners-Lee Wins Inaugural Millennium Technology Prize," Finnish Technology Award Foundation, 15 April 2004. Online at http://www.technologyawards.org/index.php?m=1&news=1&news_id=23 (checked 15 January 2007).

Supported first by the Council on Library Resources and then by the National Historical Publications and Records Commission, "Camp Pitt" brought groups largely of state archives leadership to the University of Pittsburgh campus each summer for intensive orientation on technology and program planning. Participants came away with the skills necessary to establish the first generation of electronic records programs in government archives and the knowledge that they were not alone in this endeavor.[4, 5] Many of the participants in this colloquium were veterans of Camp Pitt.

These efforts had a more limited impact on archivists outside the government or corporate arena. Archivists in "collecting repositories" have less authority to influence records creation and maintenance and limited ability to control the formats donors offer them. The disparity in circumstances among archivists has had an adverse effect on the ability of the profession to move forward on electronic records in a comprehensive and coherent fashion.

In the last decade, no doubt a result of the pervasiveness of the personal computer and the Internet, virtually all records professionals have recognized the significant impact of digital materials on how they will do their jobs. In the digital era, librarians, archivists, and records managers must be able to work with digital media as easily as they have worked with paper. They must be able to manage electronic collections, including the ability to select, acquire, describe, organize, reference, and preserve these digital works.

Managing electronic records must take at least three distinct factors into consideration. First, the rise of the Internet, especially the web, and the diminishing costs of technology has made it possible to digitize collections of "born-analog" materials. Many records managers have had to learn how to work with digitized records. For archivists who work with unique materials, digitization projects have made it possible to make their collections accessible to a much larger audience and reduce the impact of physical handling.

Second, records professionals are now working with "born-digital" materials, records and publications that may never be printed or cannot be represented in print. Given the rise of electronic information systems in business, records managers and archivists are confronted with enormous quantities of records distributed around organizations, often in decentralized systems. Many of those records can be disposed of (or discarded) before long-term preservation becomes a problem, but archivists must find ways to identify and preserve those born-digital materials that need to be kept permanently alive.

Finally, technology has always offered all professions the opportunity to work more efficiently and effectively. In the 1940s libraries began using computers to manage circulation. Later card

---

[4] Formally titled the Archival Administration in the Electronic Information Age: An Advanced Institute for Government Archivists.

[5] David J. Olson, "'Camp Pitt' and the Continuing Education of Government Archivists: 1989-1996," *American Archivist* 60 (Spring 1997): 202-214.

catalogs became OPACs, which have now evolved into integrated library systems that facilitate all aspects of library work. Archivists adopted MARC in the 1980s to ensure descriptions of their collections could be included in institutional catalogs and national union catalogs. More recently, archivists began using Encoded Archival Description (EAD),[6] an XML-based standard for finding aids to augment catalog records and meet user needs. As technology has become more pervasive, so have the opportunities to continue improving how information professionals work.

Richard Pearce-Moses, past-president of the Society of American Archivists, has argued that *what* information professionals do in the digital era remains the same.[7] They must still work with record creators and publishers to build the collection; manage the organization of the materials, their preparation for use, and their preservation; and they must work with the public and other users to provide access to the collections. None of that changes in the digital era. However, Pearce-Moses argues, *how* information professionals do their jobs will change. Many techniques used for paper records will not apply to digital works. Boxes make sense for transferring paper records organized in folders, but file transfer programs are more appropriate for digital records. Paper records can be stored on shelves in a room of stacks, but digital files will be kept on servers. Face-to-face reference service in secure reading rooms may change to asynchronous assistance for patrons accessing collections from around the world. Every activity in the physical world has a parallel in the virtual environment.

Information professionals, including archivists, have been actively engaged in designing and implementing technology in their personal and professional lives and thus are aware of the crucial issues facing the profession. However, despite the fact that most information professionals now recognize the importance of working with digital materials, many are unsure what to do.[8] The absence of rigorous standards and best practices for electronic records reflects the diverse missions of archival organizations and the use of proprietary technologies by most information creators. Many hesitate because they do not know what they need to know, and the knowledge and skill sets required by information professionals depend a great deal upon the availability of other technical expertise in the organization and willingness to collaborate across professions on their work. Do information professionals need to be able to write programs, design databases, or administer networks? Do they need to know HTML or XML, or any of a host of related acronyms (CSS, XSLT, and DTD to name a few)?

---

[6] See *Encoded Archival Description: Application Guidelines*, ver. 1.0 (Society of American Archivists, 1999) and *Encoded Archival Description: Tag Library*, ver. 2002 (Society of American Archivists, 2003). Online version at http://www.loc.gov/ead/tglib/index.html (checked 25 May 2007).

[7] "Janus in Cyberspace: Archives on the Threshold of the Digital Era," presented at the Joint Meeting of the Society of American Archivists, the National Association of Government Archives and Records Administrators, and the Council of State Archivists, Washington, DC (August 2006).

[8] See Richard Pearce-Moses, "The Perfect and the Possible: Becoming a Digital Archivist," presented at the Conference of Inter-Mountain Archivists, Ogden, UT (May 2006). Online at http://members.cox.net/pearce-moses/Papers/CIMA2006.pdf.

To help answer the question, What are the skills that information professionals must have to work with e-books, electronic records, and other digital materials?, the National Archives and Records Administration, the Arizona State Library and Archives, and the Society of American Archivists hosted "New Skills for a Digital Era." This colloquium brought together individuals with different perspectives on the question, including information professionals, educators, managers, and technologists. All were expected to have practical experience working with digital publications and records.

Discussion sessions were at the heart of the colloquium. Each session began with a presentation of one or two case studies that related to specific functions and illustrated practical skills information professionals need to work with born-digital and digitized materials, rather than merely theoretical knowledge. The colloquium sought to identify specific skills that information professionals working with digital materials needed to do their jobs. The focus was on skills that go beyond those of the consumers of records, but it seems unreasonable to expect information professionals to have the skills of a professional programmer or systems administrator.

More than sixty people attended the colloquium. The program began Wednesday evening, 31 May 2006, with a welcome by Ken Thibodeau, Director of the Electronic Records Archives Program at the National Archives and Records Administration. Richard Pearce-Moses, President of the Society of American Archivists and Director of Digital Government Information at the Arizona State Library and Archives, set the stage for discussions by framing the question. Margaret Hedstrom, professor at the University of Michigan, and Stuart McKee, National Technology Officer for Microsoft, gave keynote presentations.

During the next day and a half, the participants listened to and, in small groups, discussed eleven case studies grounded in real world experience. They assessed the skills used in the case study and suggested other skills that might be useful. Some of these skills are essential for all archivists, and others are desirable for most. Some are necessary for individuals whose specialty is working with digital materials.

Participants engaged in formal and informal conversations throughout the meeting. The case studies and keynote presentations comprised both the focus of discussion and the stimulus for wide-ranging dialogue on the broader implications of these issues. The participants were clearly energized by the level of commentary. A brave band of recorders took notes at both the general and small group discussions.

The Knowledge and Skills Inventory that follows comes directly from the discussions that took place at the meeting. This section should be viewed both as an inventory and a checklist, rather than a uniform set of requirements. The specific categories were drawn from the many sets of recorders' notes and participant commentary, and we have attempted to organize them roughly according to functional areas. Overlap was inevitable. Readers of these proceedings will need to determine the specific skills, the depth of those skills, and the knowledge areas crucial to

their institutions and to their own positions. In addition, all information professionals will need to assess the availability of expertise elsewhere within their organizations.

In his closing remarks, Archivist of the United States Allen Weinstein called for another colloquium to follow up on the work done in 2006. He recognized that the field is changing so rapidly that our understanding of the knowledge and skills we need must keep pace.

— RICHARD PEARCE-MOSES and SUSAN E. DAVIS

PREFACE

Shortly after I became Archivist of the United States, Richard Pearce-Moses, then president of the Society of American Archivists, and I began a conversation on how new technology changes the demands placed on archivists. Richard argued that *what* an information professional does remains the same. However, *how* information professionals do their jobs will change dramatically. I suggested that the greatest change would be in what the public expects of us. For example, researchers will want to see not just a selection of documents from a presidential administration, as happens today, but *all* of them. They will also want access on the Internet—immediate and total access. From those conversations came the inspiration for the colloquium, "New Skills for a Digital Era," which, in turn, led to this volume.

A relevant story: Since becoming Archivist, I have grown used to being named as one of the defendants in a number of court cases. These suits result from the increased demand for government documents, both textual and electronic. Let me provide a few examples.

When President Clinton left office, he turned over to the National Archives some 30 million electronic mail messages. A few hours after he left office, the National Archives received its first Congressional request for a subset of those e-mails. At the same time, the National Archives became responsible for a project to restore thousands of e-mails that had been "lost" by the White House document management system. That project required several years to complete and cost over $20 million.

In January 2006, President Clinton's records became subject to Freedom of Information requests under the Presidential Records Act. The requests in the first year covered more than 10 million pages of textual and electronic records. In 2007, NARA was sued because the requests for Hillary Clinton's papers while in the White House were at the end of that queue. The plaintiffs asked that the National Archives advance their request to the front of the line.

President George W. Bush will turn over to the National Archives over 100 million e-mails and nearly 100 thousand system backup tapes. Those records, no doubt, will be requested just as quickly as were the Clinton records.

It is not just presidential records that place extraordinary demands on our archivists. When John Roberts was nominated for a seat on the Supreme Court, the National Archives went to extraordinary lengths to deliver to Congress over 60,000 pages from his tenure in the Justice Department. Archivists from around the country were temporarily detailed to California to process those records, which were delivered to Congress in a remarkably short period of time. The media, of course, wanted to know why these records were not available on the Internet.

Technology has changed both the methods used to deliver information to customers as well as the scope of our enterprise. Today's cart of Hollinger boxes is a CD-ROM, and tomorrow it will

be a DVD. Today's file cabinet is the server hard drive. Our job is to collect, authenticate, and preserve valuable documents, though the definition of documents now encompasses a much wider variety of media and content.

I have great confidence that NARA's Electronic Records Archives will solve the technological challenges of preserving electronic records from relatively simple word processed documents to three-dimensional drawings from the Patent Office. I believe we have an equally daunting challenge in gaining intellectual control over these records.

The "New Skills" colloquium brought together a number of exciting presentations on how archivists are confronting the technological challenges of managing and preserving electronic records. Richard Pearce-Moses and Susan Davis have compiled these materials in a volume that will be of value to archival education for many years to come. We all owe a debt to them, to the presenters, and to the students from Arizona and Maryland who helped compile the material. Our next challenge is to develop a workshop that takes similar strides in understanding how technology demands new paradigms for managing information, and how we meet the public expectation of more information, and to have that information at one's fingertips.

ALLEN WEINSTEIN
Archivist of the United States
February 2008

KNOWLEDGE AND SKILLS INVENTORY

An old saw says, "In theory, theory and practice are the same. In practice, they seldom are." Throughout the colloquium, participants found themselves struggling with what information professionals needed to know. Although the intent of the colloquium was to consider "the practical, technical skills that all library and records professionals must have to work with e-books, electronic records, and other digital materials,"[9] the discussion often touched on basic knowledge practitioners needed to be able to work wisely.

The discussions at the colloquium illustrated that professional knowledge relating to electronic records and publications takes many forms, drawn in part from the larger environment and in part from the specific context for the field. We refer to the larger environment as the "information ecosystem," borrowing from Thomas Davenport's *Information Ecology*.[10] Davenport describes the whole of an organization's information environment—how information is used, both to accomplish work and as a tool of power, and the way that information is organized.

KNOWLEDGE

The topics described below do not comprise all the knowledge that information professionals need. Rather they provide a framework for the new or expanded knowledge necessary to thrive in the digital era. Armed with this knowledge, information professionals can be a vital force in transforming the information ecosystem. Some areas are familiar ground; others reflect the collaborative environment in which information professionals operate.

THE INFORMATION ECOSYSTEM

Information takes many forms and is used for many interrelated purposes. A variety of forces influence this complex system. In some ways, the roots of the current information ecosystem lie in the Industrial Revolution and the rise of systematic management.[11] If information professionals expect libraries and records programs to be a dynamic, thriving part of the information ecosystem, they must understand their environment. Not only do they need to comprehend the

[9] New Skills for a Digital Era [home page], online at http://rpm.lib.az.us/newskills/ (checked 25 May 2007).

[10] Thomas H. Davenport, with Laurence Prusak, *Information Ecology: Mastering the Information and Knowledge Environment* (New York and Oxford: Oxford University Press, 1997). According to Davenport, information ecology "emphasizes an organization's entire information environment. It addresses all of a firm's values and beliefs about information (culture); how people actually use the information and what they do with it (behavior and work processes); the pitfalls that can interfere with information sharing (politics); and what information systems are already in place (yes, finally, technology)."

[11] See JoAnne Yates, *Control Through Communication: The Rise of System in American Management* (Baltimore: Johns Hopkins University Press, 1989).

information ecosystem at the macro level, they must also understand variations in their local environment. If they isolate themselves in a backwater, they—and their programs—will stagnate.

One of the most important things to understand is how technology has transformed the information ecosystem. If telegraphy formed much of the modern recordkeeping systems in the nineteenth century, the Internet has certainly changed the context of publications and records. Information is now widely available, and access continues to expand through the Internet, PDAs, and mobile phones. Not only is information ubiquitous, the quantity of information continues to grow exponentially, and the speed of information transfer continues to accelerate.[12]

The metaphor of an ecosystem is particularly appropriate to records managers and archivists, who have long spoken of the information lifecycle. Archivists have traditionally emphasized the importance of understanding the context in which records are created and used by records creators. Thus it follows that archivists should be equally concerned with both utilizing and documenting this information ecosystem.

During the colloquium, discussion touched on a number of themes that expand on what information professionals need to know about the information ecosystem.

Information Architecture

Participants often used the term "information architecture" to describe what other information professionals mean when talking about the information ecosystem.[13] Where "ecosystem" carries connotations of something organic that evolved over time, "architecture" connotes something synthetic and intentionally designed. This subtle shift in meaning suggests that information professionals must not only understand the world they live in, but how to manipulate that world. They must know how the pieces can be fit together to build something useful. The participants also used the phrase "information management" with similar meaning.

Standards

In order for different systems to work together in a technological environment, information professionals need shared specifications for interoperability. In the information ecosystem, that means understanding a wide range of standards for different purposes. For example, metadata standards have been established for administration, description and discovery, and preservation of information. Numerous standards relate to packaging information and to exchanging information among systems. Here we note that the participants felt it was important for infor-

[12] Yates.

[13] "The structure and interrelationship of information, especially with an eye toward using business rules, observed user behaviors, and effective interface design to facilitate access to the information." *A Glossary of Archival and Records Terminology* (Chicago: Society of American Archivists, 2005). "Information architecture" should be distinguished from "systems architecture" and "information systems", which refer to the implementation of a specific architectural design in hardware, software, and procedures.

mation professionals to be aware of established and developing standards, and they discussed how hard it was to know which of those would have an impact on how they did their jobs.

The Open Archival Information System (OAIS)[14] is a reference model rather than a formal standard. It serves as a framework around which a coherent set of standards can be developed to address the many different issues associated with this complex issue. As such, it can be considered a good model for the information ecosystem.

Legal Context

Laws are a specific type of standard, codified by statute, regulation, and judicial opinion. Records managers and archivists have always understood that the law has a direct impact on which records needed to be kept and for how long, and librarians have become very sensitive to the retention and use of patron records in light of the Patriot Act. However, recent events have made the legal aspects of recordkeeping more visible. No doubt the Enron debacle and the Sarbanes-Oxley Act have made compliance a hot topic in corporate America. Further, recent changes in the Federal Rules of Civil Procedure have a significant impact on preservation and discovery of electronic records.[15] Many participants saw the increased awareness of compliance and risk of litigation as a wave that information professionals could ride to greater influence within the organization. The rise of digital rights management, as well as recent and potential revisions to copyright law, influence the reproduction of materials, including provision of online access to collections. Finally, recent news-making events regarding identity theft raise concerns for protecting individuals' privacy, especially when mounting archival and biographical information on the web.

Trend Spotting

One important challenge facing information professionals is keeping up with the rapid pace of change. Participants believed that the professions need to understand trend spotting—sometimes called horizon scanning—so that they can anticipate and plan for changes in the information ecosystem.

Ethnography and Anthropology

The information ecosystem includes more than information. It is inhabited by creators, distributors, consumers, and custodians of information. Information professionals, whose mission is to select information to be stored and to redistribute that information to future consumers, must know something about the other "natives" in the information ecosystem to be able to work effectively and collaboratively.

---

[14] *Reference Model for an Open Archival Information System (OAIS)* (Consultative Committee for Space Data Systems, 2002), online at http://public.ccsds.org/publications/archive/650x0b1.pdf (checked 25 May 2007).

[15] See "New E-Discovery Rule Amendments Proposed," *The Third Branch* 36, no. 7 (July 2004). Online at http://www.uscourts.gov/ttb/july04ttb/ediscovery/index.html (checked 10 February 2007).

Creators are of particular interest. Records managers are primarily concerned with creators who function within an institutional context. Archivists are interested in those same creators, but those in collecting archives are also interested in creators working individually. Librarians are also interested in both. Information professionals must understand creators' motives; are they creating information for profit (motion pictures, music, novels), public service, or vanity? Do they charge for the information, derive revenue through advertising or other streams, or do they give it away for free? Do they work with distributors or publishers, who expect some sort of remuneration?

Consumers are another key group within the information ecosystem. What do they want and need? What are they willing to pay for? How do their needs change over time? When do they need current information? Historical information? When do they need the raw data in records versus the synthesized data in publications? What are their concerns for the trustworthiness of the data? While seeking answers to these questions, information professionals are aware that collecting information about patrons and other users may conflict with privacy issues noted earlier.

Information technologists are a critical group within the digital information ecosystem. They support all aspects of the use of information. Several participants pointed out that many references to information technologists were naive. They are often treated as a monolithic group, when in fact there are as many types of technologists as there are archivists, librarians, and records managers. Information professionals must be able to distinguish network, system, and database administrators. They must know if they need the skills of a systems analyst or programmer. Someone really good at desktop support may be useless as a web designer. No one information technologist has all technical skills.

Understanding this universe requires skills of observation that information professionals can learn from anthropologists and ethnographers. Anthropologists examine humanity through physical, social, economic, political, and cultural characteristics over time. They, and other researchers, use ethnography as a fieldwork-based method for describing human social phenomena. This approach can shed light on the interrelationships among individuals and organizations, and the issues and events underlying those interrelationships.

INFORMATION STUDIES

If one element of our knowledge base is information ecology, another is information studies.[16] Not only do information professionals need to understand the larger context of their work, they need to be masters of their own disciplines. Records managers, archivists, and librarians offer specific perspectives on the information ecosystem. As stated earlier, archivists and other information professionals have struggled to determine the relevance of the tried-and-true

[16] We are using this term to incorporate archival studies, library and information studies, and records management, recognizing that each of the information professions takes a different approach to education.

principles of their work in an electronic world. The colloquium reaffirmed that it is essential to have a strong core grounding in information studies. These disciplinary principles allow us to apply skills wisely.

While the core archival principles and functions remain, practice is changing. This shift means that information professionals must reconceptualize many principles that serve as the foundation of traditional practice. Respect for provenance is a core archival principle, used to facilitate access and to demonstrate authenticity. However, the ease with which records can be copied and distributed, coupled with the fluidity of organizations, makes it difficult to ensure that records from one source have not been mixed with those of another. Archivists have relied on original order to preserve context and facilitate description, but that model—based on a specific physical arrangement—is meaningless in a technical environment where creators may retrieve and sort records using different queries returning different results. What is the meaning of "original" in an environment where the information is virtual and can exist in multiple, identical copies, each of which can be authenticated?[17]

Even the very question of libraries and archives taking custody of records is up for debate.[18] The ways in which we carry out functions of acquisition, appraisal, description, storage, preservation, and access are all evolving.

Information professionals must also reexamine their assumptions about the materials that fall within their responsibility. Many of the core professional concepts pertain to the content of records, and the ways in which information professionals manage and maintain that content. Some concepts relate more to the carriers of that content. In an analog world practitioners are concerned with fairly basic format issues: paper and film, single sheet and bound volume, oversized items. In a digital world a number of format issues (described below) remain important because of the wide range of possibilities and the complex interrelationships among software, hardware, and storage media. These issues raise concerns about many of the core concepts on which information professionals have relied.

Blurring boundaries between publications and records. All three information professions inherit a long-standing debate as to what is considered a publication, and what is considered a record. Possibly one of the most important things that an information professional must be able to do in the digital era is to understand how the terms *publication, record, document,* and *information* overlap, how they are used within different communities, and where differences in terminology are relevant. The ultimate goal is to ensure that useful information is captured in a form that can be managed by the repository.

---

[17] See David M. Levy, "Where's Waldo? Reflections on Copies and Authenticity in a Digital Environment," in *Authenticity in a Digital Environment* (Council on Library and Information Resources, 2000). Online at http://www.clir.org/pubs/reports/pub92/levy.html.

[18] David A. Wallace, "Custodial Theory and Practices in the Electronic Environment," *SASA* Newsletter 2002:1 (January-March 2002). Online at http://www.archives.org.za/nletter/nletter2002-1.PDF.

Context of creation and use. In archival terms, the ways in which records were created and used by the records creators is a critical factor in determining long-term value. In an analog world, records were usually fixed in form at the time of creation. In a digital world, creation and use is a fluid process, one that is more difficult to define and evaluate.

Life cycle approach. In records management, records pass through a series of stages from creation to disposition. Archivists may identify records for permanent retention at an early stage, or they may have no early role in the appraisal process. Transfer of custody takes place when records are no longer in use by the records creators. In a digital world, custody may never shift, and without archival input, important records may not survive the active stages of maintenance and use. It becomes crucial for archivists to be involved when records systems are created and maintenance decisions are made.

Framework for arrangement. Archivists rely on core concepts of provenance and original order to make decisions on arrangement. Description of materials is based on intellectual and physical aggregations that result from arrangement decisions. Arrangement of digital records cannot occur in the same fashion, perhaps bypassing the question of original order. Even provenance is less secure when records creation is shared in an electronic system.

Integrity of records. Integrity involves both authenticity and reliability. The SAA Glossary defines authenticity as "the quality of being genuine, not counterfeit, and free from tampering, and is typically inferred from internal and external evidence, including its physical characteristics, structure, content, and context." Reliability reinforces the concept of the records' trustworthiness. These attributes are more difficult to ensure in a digital world, where shared systems allow easy manipulation of records. In addition, maintaining chain of custody is often impossible, increasing threats to the integrity of records.

Intellectual control. Information professionals have evolved descriptive practices that take advantage of automated techniques for bibliographic control. The advent of OPACs and other online systems (e.g., hierarchical, aggregate descriptions; standards) have led archivists to adopt standards, many of which come from the library world. Archival description is based on hierarchical levels of control and aggregate, not item-level, description. In a digital world, hierarchies often disappear, and item-level metadata predominates. Intellectual control thus follows a very different course.

Physical control. In an analog world, archivists separate intellectual control from physical control. Analog records can only be in one place, in one filing system, while finding aids can refer to records in a more flexible fashion. Archivists bring together intellectually items that are physically separate. In a digital world, physical control relates more to issues of preservation and storage, as records lack the same physical presence as they did before.

Inferring relationships between digital and analog formats. Information professionals have always dealt with issues of reformatting. Textual records were microfilmed; fragile items were photo-

copied; photographs had both negatives and prints. In a digital world, records may exist only in electronic form, or they may represent the digitization of analog materials for purposes of preservation or increased access. The same record series might exist on paper for one time span, and electronically in another. Information professionals must develop systems that are flexible enough to encompass a range of possibilities.

Post-custodialism. Some archivists question if archives will have the skills and resources to keep an extraordinarily diverse collection of electronic records that rely on many different combinations of hardware and software. Post-custodialism questions the assumption that records will be transferred to the archives, suggesting that archivists do not need technical skills to support records maintained on external systems.

Best practices. In all of these issues, information professionals need to develop best practices. What differs in the digital world is the speed at which these practices must evolve and the necessity for collaboration and communication.

DOCUMENTARY FORMS IN THE VIRTUAL ENVIRONMENT

Possibly the most significant driving factor in the transformation of the information professions is the simple fact that the very stuff of the professions is changing from analog to digital signals, from physical to virtual formats. Although this transformation began long ago—Hollerith cards were developed in the nineteenth century, as were Edison wax cylinders and motion pictures—the majority of information continued to be produced on paper until the late twentieth century.

This shift from paper and analog formats means that information professionals must learn about the nature of the materials used to store information. In some cases, the virtual format parallels the physical; for example, a word processed letter is virtually identical in its intellectual form to its paper predecessor. However, information professionals must also learn about entirely new forms. For example, a geographic information system is more than a set of maps; it is a database that can respond to queries based on space and time. As Peter Wilkerson noted, "We don't just want fast paper, but analyzable data."[19]

In the virtual world, there may be no "record" as traditionally understood. Some systems may not fix information, with the result that the information is constantly changing. Unless a system is designed to fix information so that someone can be assured it has not been altered over time, the system will contain no reliable record of the past. For example, a geographic information system may not store historical data; unless the system can roll back updates, it cannot show a previous view. When information is constantly changing, records professionals may want to work with records creators to capture periodic snapshots of the database. If keeping the software necessary to render the information is prohibitively complex or expensive, it may be

[19] Recorder's notes from the colloquium.

possible to export selected reports or datasets to fixed information and software-neutral datasets for future analysis.

Throughout the discussion, participants stressed the importance of understanding the essential attributes of a record. In the digital era, information professionals need a rich understanding of the theory of recordkeeping and publication so that they understand the *why* as much as the *how* of their work. As one participant noted "We have to know what's okay to lose; we have to know the essence of what we can and want to keep, including the data, metadata, format, and moment of recording."

Participants mentioned many different intellectual and virtual formats with which information professionals should be familiar, although—surprisingly—there was limited discussion of e-mail outside Dymalski and Kirkpatrick's case study.[20] Participants did not have the time to determine which formats *all* information professionals should understand, as opposed to those formats that are appropriate for specialization. In many ways, this discussion drove the larger conclusion that not all information professionals will need the same skills, nor will all need the same level of expertise.

Data Formats and File Types

Digital information can be organized into any number of high-level categories, which are not mutually exclusive. The information may be in character or binary format, the former human readable and the later meaningful only to machines. Either format may be used for similar purposes. WordStar, for example, stored the documents in character format, and the text, interspersed with occasional codes used for formatting, could be read using a simple ASCII text editor. Microsoft Word stored the text and formatting of documents in a binary code; opening the file in a simple editor revealed a seemingly random sequence of bytes, although the forthcoming version of Word will store documents and formatting as character data.

Data formats can be categorized in terms of the file types used to create or store the information. Participants felt that information professionals should at least be able to identify the most common file types, such as .doc, .xls, .mdb, and .ppt (Microsoft Word, Excel, Access, and PowerPoint); .rtf (Rich Text Format); .wpd (WordPerfect); and .pdf (Adobe Acrobat). Image files may be .jpg, .jp2, .jpx, .gif, .tiff or .tif, and sound files include .avi and .wav. Database file types include .mdb (Access), .dbf with .dbm (dBase), and .odb (Open Database Format).

The more information professionals understand about the manner in which documents are encoded, the better they will be able to ensure that those documents will be preserved in a manner that protects their authenticity and integrity. The participants felt that all archivists should understand common file types, such as those mentioned above, though archivists who work closely with any specific format will likely need more sophisticated knowledge of specialized

[20] "One County's Attempt to Move from 0 to Digital in Record Time."

file types relevant to that format. In particular, information professionals should know when to recommend one format over another for long-term preservation.

Databases

One particular category of data storage deserves particular attention because of its importance, widespread use, and complexity. Databases contain "information that is accessed and updated through software (a database management system) that has been organized, structured, and stored so that it can be manipulated and extracted for various purposes."[21]

Relational databases store information in flat files, each row a separate record and each column used for a distinct field in the record. The columns may be fixed width or delimited by a special character. The data may be stored in a number of related tables or in a sequential string, using tags, delimiters, and an internal directory to identify fields. Other database technologies (for example, ISAM, XML, object-oriented systems) store the information differently. The fact that virtually every database organizes the information in a different manner means that information professionals must not only be familiar with the general concept of a database, they must be able to understand something of the internal structure of the data. In many instances, a database will have accompanying documentation, although the database must be validated against that documentation to ensure it is accurate. In a worst case scenario, when an important database has no documentation, an information professional must consider digital archaeology to excavate and reconstruct information about the database.

Information professionals must understand the distinction between dynamic information and records, which by definition contain information that has been fixed in time. Although databases may contain predefined views of the data (records) and reports (selections of records), they can be queried in many different ways. The ability to generate records and reports different from those used by the creator of the database offers great opportunity for further analysis, but also raises important questions about the authenticity of what a patron is viewing: is it a view created in the course of business (a record) or something created later?

Markup Languages

Discussions of data formats may become less significant as the technology industry seems to be moving toward a single standard, Extensible Markup Language (XML). XML typically stores textual information as tagged character data whenever possible, although non-textual data, such as images or sound, are stored as tagged binary data. Participants believe that XML is becoming the *lingua franca* of the digital era and that all information professionals need at least a rudimentary knowledge of that standard. It is possible that the majority of information professionals will need to be fluent in basic XML, as well as related standards, such as Extensible Schema Description (XSD), and Extensible Stylesheet Language Transforms (XSLT), to name a few.

[21] *A Glossary of Archival and Records Terminology*.

Information professionals may also need to know something of other markup languages, such as Hypertext Markup Language (HTML), now largely a subset of XML, and Standard Generalized Markup Language (SGML).

Media

Floppy disks, now largely obsolete, can only be read using hard-to-find drives. Compact discs will soon meet the same fate. Tape comes in dozens of formats, and the manner in which information has been recorded on the tape depends on a number of factors, including the hardware and operating system. Information professionals must be familiar with the wide range of media used to store digital information. Information professionals must consider if the meaning, significance, or value of the media that stores the original bitstream. Are there physical labels that need to be preserved for context? Is there intrinsic value in the media? (See also Preservation, below.)

SKILLS

The skills that emerged from discussions at the colloquium fall into three broad categories: management skills, technical skills, and soft skills. These skills are framed within the context of relevant professional knowledge. They are tools which information professionals use based on the theoretical knowledge specific to their profession. Without that guiding knowledge, an individual acts as a technician rather than a professional.

MANAGEMENT SKILLS

Information professionals need a wide range of skills to administer the ongoing operations of a successful organization, ranging from fiscal to facilities, from planning to evaluation. This is an area that engendered considerable discussion. If managers are overseeing, rather than directly carrying out specific tasks, how much technical knowledge do they need to have? Is it enough to have someone on staff who possesses that knowledge? The consensus was that, in most cases, managers do not need the level of technical expertise required for specialized staff. However, managers cannot absolve themselves from responsibility; they need to understand broad concepts and applications in order to plan and evaluate effectively. All information professionals must take responsibility to learn enough about digital information to do their jobs.

In the digital era information professionals must be familiar with the technology in their holdings, but they should also be able to use technology to do their jobs more effectively. Card catalogs and indexes have given way to online catalogs and databases; paper forms are now data entry screens. Finding aids are word processed, rather than typed, and preferably marked up using EAD.

Ultimately, information professionals must be able to reengineer their own business processes and recordkeeping to take full advantage of the benefits of technology. To do that, they need to be able to apply the same skills to their own work, rather than to their collections. Automating curatorial activities and workflows within the repository is a complex task. Participants recognized that information professionals are not expected to be professional technologists. Even so, they felt that information professionals needed sufficient understanding of the process to be able manage that process effectively and to ensure that programmers (on staff or contractors) were making wise choices. Participants often expressed the need to talk to other communities, and the more information professionals are familiar with systems design, the easier it will be for them to articulate the repository's needs.

Information professionals, especially those in management, will continue to need administration, project management, and evaluation skills. Participants identified a number of skills in these areas that require technical knowledge or skills for success.

Information professionals will continue to develop budgets, as well as plan and prioritize work, but given the high costs and complexity of information systems, the ability to do a cost-benefit analysis is critically important. Similarly, many information professionals negotiate contracts, but they need to write clear specifications for the work and determine if a vendor's proposal is reasonable and makes financial sense. After the contract is begun, information professionals must be able to evaluate performance using both qualitative and quantitative analysis for effective quality control. To get records creators to comply with records management or archival programs at the organizational level, information professionals must be able to write effective policies and procedures.

Process may be one of the most challenging areas for information professionals. Helping a repository and its staff move from working with physical formats to digital formats requires a number of new skills. Many information professionals have never gone through such a significant change in the way they work. Participants suggested a number of skills that can help, including project management, change management, business process reengineering and process improvement, and workflows analysis. One key skill is managing expectations, of staff and patrons, as well as management.

TECHNICAL SKILLS

Technical skills are those requiring educational or experiential grounding in that function or activity. This contrasts with management and soft skills, which rely more on the ability to communicate and coordinate those activities. For most of the colloquium participants, technical skills were broadly defined, ranging from computer-related activities to more traditional archival functions of appraisal and selection. Clearly, no single person will possess equivalent expertise in all areas. However, within a repository, all areas require attention.

For example, photograph archivists commonly write a unique identification number on prints; similarly, digital archivists may assign a unique filename to electronic records. When working with thousands of files, this task could be done much more quickly using a simple script than manually renaming files.

The ability to develop a home-grown, integrated system that automates every activity within a repository is certainly beyond anything an information professional is expected to be able to do. However, many activities produce relatively simple records that are perfectly suited to a desktop database system such as Microsoft Access or OpenOffice Base. While many participants felt information professionals do not need this level of IT skills, virtually all felt that the ability to take full advantage of desktop office tools (word processors, spreadsheets, databases) would enable information professionals to work more efficiently. Basic knowledge of desktop office tools can serve as a foundation for working with professional programmers.

The participants' discussion of the need for information professionals to have programming and systems skills was very spirited. More than a few felt that it was important for information professionals to have these skills,[22] while others—possibly the majority—disagreed.

Ultimately, the group came to the realization that not all information professionals will need the same knowledge or the same level of expertise in technical skills. Managers need to know enough to plan and supervise projects; for example, they should understand how moving to an open source application could require additional IT staffing to maintain the software, and they need to be able to assess whether a program built around a specific application can be sustained. Practitioners working directly with digital materials will likely need at least a rudimentary knowledge of programming (the ability to write simple scripts, if not at a level necessary to develop large-scale applications). For example, sorting and filtering a list of files in a directory, searching for specific text strings in a large group of files, moving files from one location to another, or working with records in a database can be done with a few lines of code in Perl, Java, or Visual Basic. If information professionals can do simple tasks using simple scripts, they need not wait for a professional programmer.

[22] For example, Pearce-Moses' belief that archivists must acquire IT skills necessary to adapt to the digital era was a driving force behind this colloquium. See "A Bridge to the Future," keynote at the Society of Southwest Archivists, May 2005 (online at http://www.lib.az.us/diggovt/presentations/Bridge.pdf, checked 18 February 2007). "The Winds of Change: Blown to Bits," incoming presidential address, Society of American Archivists, 2005 (online at http://members.cox.net/pearce-moses/Papers/WindsOfChange2005.pdf, checked 18 February 2007). "Janus in Cyberspace: Archives on the Threshold of the Digital Era," presented at the Joint Meeting of the Society of American Archivists, the National Association of Government Archives and Records Administrators, and the Council of State Archivists in Washington, DC, August 2006 (online at http://members.cox.net/pearce-moses/Papers/Janus_abridged .pdf, checked 18 February 2007).

RECORDS MANAGEMENT

Records professionals are more actively involved with the creation of records than librarians, though the blurring of professional boundaries diminishes this difference. Although records managers and archivists share the goal of protecting information of enduring value, records managers focus on records during the active and inactive life while archivists focus on those records that have been set aside for permanent preservation. Furthermore, records managers typically work within an institutional context, while archivists may work with personal papers as well as institutional records.[23] Frequently the same individual wears both hats.

Many of the skills necessary for a successful records management program relate to the ability to work well with people—the individuals who create and manage records during their active and inactive life. Although these are not technical skills, they surfaced frequently during the colloquium, often enough to warrant separate discussion later in this report.

Participants identified a number of skills that should take place before destruction or transfer to an archives.

Systems Analysis and Design[24]

Information professionals must understand any activity or workflow in order to successfully automate it. One of the first lessons of automating a project is discovering the inconsistencies, assumptions, and undocumented steps in a manual system.

Participants felt system analysis was essential to enable records professionals to be involved in the specifications for an electronic recordkeeping system, whether that system would be purchased or custom built. Failure to have a records professional at the table could mean that critical recordkeeping concepts might be overlooked, resulting in a system that did not have trustworthy records (such as "records" with content that can be inadvertently changed), that could not perform even basic records management functions (such as deleting records), or could not export records into a future system or a separate electronic archives.

Often records creators want advice on specific document management systems, also described as content management, version control, imaging, or e-mail management systems. Systems analysis skills will help records professionals provide guidance.

Participants felt that system design[25] skills fall outside the realm of information professionals, although familiarity with those skills is desirable.

---

[23] Needless to say, this statement greatly oversimplifies the two professions.

[24] "The analysis of the role of a proposed system and the identification of a set of requirements that the system should meet, and thus the starting point for system design." *Dictionary of Computing*, 4th ed. (Oxford University Press, 1996), 491.

[25] "The activity of proceeding from an identified set of requirements for a system to a design that meets those requirements. A distinction is sometimes drawn between *high-level* or *architectural design*, which is concerned with the

Business Process Reengineering[26]

An automation project can be an excellent opportunity to improve and modernize manual systems. To accomplish such a project, information professionals should be able to take a holistic view of a repository's activities, and develop a smooth workflow that eliminates redundant steps and increases efficiency and effectiveness. At the same time, this knowledge will enable information professionals to participate in larger business process reengineering in the organization to ensure that their needs are met.

Modeling and Prototyping

The ability to model an abstract process, such as accessioning or preservation migration, through a description or diagram, is a valuable skill for system analysis and business process reengineering. Prototyping a model to test and refine it is similarly valuable.

Classification and Metadata

Participants noted that a system designed to support records management and archives greatly improves the quality of those programs. Classification and metadata are two key elements of such a system. Records that are properly classified in groups can be treated as aggregates. Records with appropriate metadata can be managed and retrieved much more effectively, and can be migrated and checked for integrity more easily. Participants noted information professionals must be experts in these two areas.

Advocacy and Outreach

To be able to participate effectively in these discussions, records professionals should be able to communicate records requirements and to explain how those requirements fit into the larger system. Although advocacy and outreach are generally "soft" skills (and will be discussed later in that section), getting support for a records management program is so dependent on these skills that they are mentioned here as well.

Specific skills involve identifying key people, finding the right message for different audiences, and negotiating policies. Participants also pointed to change management[27] as a useful skill. In these cases, records professionals should be prepared to address issues of increased efficiency and effectiveness, pointing to how reengineering the recordkeeping process can save time and money by eliminating duplicate information. The better information professionals can help re-

---

main components of the system and their roles and interrelationships, and *detailed design*, which is concerned with the internal structure and operation of individual components." *Dictionary of Computing*, 490.

[26] "Redesigning the way activities in an organization (business processes) are carried out to improve efficiency and reduce costs." *A Glossary of Archival and Records Terminology*.

[27] "Planned, systematic alterations to established missions, objectives, policies, tasks, or procedures within an organization. Change management typically refers to an intentional process undertaken by management in response to internal needs. However, it may also include strategies for responding to external events." *A Glossary of Archival and Records Terminology*.

cords creators adapt to a new workflow that supports records management, the more likely the success of the records program.

In some contexts, risk management and legal compliance may be the key, and emphasizing the value of records to protect the organization in case of litigation or audit is often an effective approach. Several participants pointed to the University of Pittsburgh's project to identify warrant for recordkeeping as core knowledge to bring to these discussions.[28] Other audiences may find such a message threatening.

Managing Expectations

At the same time, records professionals must be able to manage expectations. Many records creators look for a simple solution, and records professional must be able to explain what technology can and cannot do. For example, many people want a single retention period for all e-mail, failing to recognize that the content of e-mail is the driving factor, and that litigation holds further complicate the issue.

SELECTION AND APPRAISAL

Fundamental questions face information professionals regarding how the value of information has changed in the digital era. Has the transformation from paper to digital formats altered the content in ways that affect the records' value? How has the vastly increased quantity of information changed the appraisal process? Given the speed at which records are created, many may contain errors that affect their reliability, and the rapid proliferation of documents raises issues of duplication and version control.

Participants noted that information professionals will remain actively involved in selection and appraisal. These tasks cannot be delegated to records creators. While records creators may be able to play a role, information professionals are ultimately responsible for the collections and, if nothing else, must audit what records creators are transferring to the repositories. One participant used institutional repositories as an example where the model of records creators archiving their material was not nearly as successful as had been hoped.

Functional Analysis

Participants considered functional analysis, "a technique that sets priorities for appraising and processing materials of an office based on the relative importance of the functions the office performs in an organization,"[29] one approach that could be useful in appraising massive quantities of records. In many ways, functional analysis is a strategic approach to collecting (although in some contexts, *triage* may be a more appropriate metaphor). Given the massive quantities of information and the limited resources to appraise those records, records professionals must

[28] See Wendy Duff, "Harnessing the Power of Warrant," *American Archivist* 61, no. 1 (Spring 1998): 88–105.

[29] *A Glossary of Archival and Records Terminology*.

concentrate on the most important, with the understanding that less important records may be lost. This approach assumes that "importance" correlates with organizational hierarchy and key functions. Thus, in a state context, emphasis would be put on a governor's records before those of the barbers' commission, or within the department of administration, policies before housekeeping documents.

Computer Assisted Appraisal

It is surprising that there was little discussion of the potential use of artificial intelligence at the colloquium, even though some commercial electronic recordkeeping and document management systems offer this function. A few participants mentioned the use of natural language processors and rules-based systems, which might be seen as an indirect reference to the use of AI. These ideas may not have come up because only a few attendees had experience using these tools, suggesting that this area cries out for research to adapt skills from computational linguistics and the legal community[30] to records management and archives.

Surveying Electronic Records

When working with paper records, records managers and archivists often surveyed records by going into file rooms, offices, basements, and worse environments, clipboard in hand, making a visual inspection. The equivalent activity in the digital era might be to log into systems and traverse the file system. At least some participants felt knowledge of operating systems and their file systems would be a useful skill. However, others felt such an approach would have only limited success in the digital era. Few systems administrators would give records professionals unlimited access to a system so that they can poke around. Moreover, the file names may not be a good indication of the content. And the size and complexity of file systems, even in smaller organizations, make this impractical with the use of specialized software.

One of the best approaches may be interviewing the individuals who create and use records. Here, business process analysis is a supporting skill, giving records professionals the ability to ask the right questions. At the same time, participants recognized that this approach also had its limitations; in medium to large organizations, it's unlikely that records managers could meet all the record creators. Rather, they would have to set priorities, and macro appraisal and functional analysis could help identify those it was most critical to interview.

Scheduling and Records Disposition

In the end, appraisal decisions must be documented in some fashion. In an institutional environment, record schedules are commonly used to indicate whether routine records should be destroyed or transferred to the archives. Participants noted that record schedules can help raise awareness of retention of digital information throughout the organization. Participants also

[30] For example, the Sedona Conference has developed a draft document describing the use of search engines in document discovery and production. "Best Practices Commentary on the Use of Search and Information Retrieval Methods in E-Discovery." Private communication with the author; the report has not yet been released to the public.

noted challenges related to scheduling and records disposition with databases and other digital formats. Participants suggested that records professionals must know techniques for the secure destruction of records and data to protect confidential information, the digital equivalent of shredding and macerating.

ACQUISITION

Information professionals will continue to build collections. In the digital era, those collections may include virtual holdings, documents that are not held by the repository but represented through links to external resources. Ideally, archives will take possession of the records so that they can ensure their preservation in a professionally managed, standards-based environment with a sustainable resource base. In reality, some archives will not have the resources to do that and may rely on agreements with record creators or others to store the records. The post-custodial theory of archives suggests one approach, while trusted digital repositories suggest another.[31, 32]

Digitization

Participants discussed the role of digitization in the digital era. Digital archives may acquire this class of electronic records when record creators use digital imaging rather than microfilm. Similarly, repositories may digitize their collections, although participants clearly distinguished the use of this technology for increased access from the use of digital imaging as a preservation technique.

Some participants felt digitization was out of scope, wanting to focus on born-digital records; once a record was digitized, it wasn't significantly different from other digital records. However, the majority felt that for the foreseeable future information professionals would need to understand digitization. Many record creators digitize paper records, and information professionals must be able to give the record creators guidance on best practices and quality control to ensure that the information can still be accessed after digitization. Information professionals may also digitize their analog holdings and must apply the same best practices and quality control mechanisms.

File Transfer

Beyond the theory of what to acquire, records professionals must consider the very mechanics of acquisition. Archivists may use a variety of technologies to transfer digital records. In some instances, they may use external media such as portable hard drives (for limited quantities of data, more appropriate to an individual's personal records) or tape. For large systems, they

---

[31] David Bearman and Margaret Hedstrom, "Reinventing Archives for Electronic Records: Alternative Service Delivery Options," *Electronic Records Management Program Strategies*, Archives and Museum Informatics Report 18, ed. Margaret Hedstrom (Pittsburgh: Archives and Museum Informatics, 1993), 82-98.

[32] "Trusted Digital Repositories: Attributes and Responsibilities: An RLG-OCLC Report" (RLG, May 2002). Online at http://www.rlg.org/legacy/longterm/repositories.pdf (checked 12 February 2007).

must know a variety of protocols to transmit files over the Internet, such as ftp or rcp. Further, information professionals need a firm understanding of networking and security best practices to know how to negotiate openings in firewalls and protect confidential records from hackers.

Validation

In order to be able to demonstrate the authenticity of records, information professionals need to certify their integrity. Were the records unintentionally altered during the transfer process? Were all the records received? According to Linda Henry of the Center for Electronic Records at the National Archives and Records Administration, validation (also called verification) includes "comparing the electronic records . . . received with the description of the records that the agency provided." Referencing her article with Tom Southerly, she continues, "After acquiring data from a federal agency, the Center processes include preservation copying and verification of data files before making them available for use. The verification portion of this process does not assess the quality of the data. Rather, it is the process of comparing the content of records received from a federal agency to the description of those records as represented by the record layout and codes provided by that agency during the transfer of those records. Because archivists do not do analytical research but preserve records for use by others, they don't rely upon verification by widespread research use."[33] In the context of XML records, validation includes verifying that the document is well formed (the structure of the document is correct) and complies with a document type definition or schema. Validation may take place as part of acquisition or later in processing.

Middleware

Middleware is designed to automate business processes by implementing rules in software. Acquisition is a perfect opportunity for middleware applications as it is designed to mediate "the exchange of information between two applications or between an application and a network,"[34] in this case, an exchange from a record creator's system to the archives' system, typically via a network. For example, the acquisition workflow might begin when an agency deposits records on the repository's ftp server. The middleware then picks up the records, transforms them from their native format to one designed for the long-term storage of the records, assigns administrative, discovery, and preservation metadata using a variety of routines, creates entries for the records in an accessions register, and then deposits the records in the "digital stacks." Middleware may be used to validate records.

Software vendors often sell middleware as a tool that "knowledge workers" can use without significant knowledge of the underlying technology. Many packages have a graphical user interface intended to hide the code that drives the system. In fact, configuring middleware to

---

[33] Private correspondence with Linda Henry of the Center for Electronic Records at the National Archives and Records Administration, 26 January 2007. She cites her article with Tom Southerly, "Archivists and Statistical Literacy" in *Of Significance...A Topical Journal of the Association of Public Data Users: Statistical Literacy* 1, no. 1 (1999),: 31-34.

[34] *A Glossary of Archival and Records Terminology.* IBM's WebSphere and Microsoft's BizTalk are examples of middleware applications.

facilitate transfer, processing, and storage of electronic publications and records will likely require some programming knowledge, especially when the workflow includes fairly advanced transformations to prepare the record for long-term storage and analyze the record content to create discovery metadata. Larger repositories may be able to hire a programmer with middleware skills, but information professionals in smaller shops may need to learn enough programming (often relatively simple scripting) to take advantage of middleware tools.

Harvesting Web Publications

Middleware assumes a negotiated transaction between the records creator and the repository. However, libraries and archives that want to collect publications from the web need another model. Acquiring web publications is an informal convention; publishers post the information, and it is up to the repository to get it.[35] A number of libraries and archives are using spider software such as wget and Heritrix to harvest websites to acquire information posted on websites.[36] As noted above, information professionals are not necessarily expected to have the skills to write programs that use those spiders, but they must be able to develop the software specifications, use and configure the resulting application, and evaluate whether the application is operating as anticipated.

PROCESSING

The boundaries between acquisition, processing, and subsequent functions are often blurred when working with digital records. For example, Margaret Adams noted "Because digital records are not tangible objects, informed reference service is totally dependent upon the intellectual control and other products of archival processing."[37]

---

[35] The practice of harvesting web publications provides a good example of the complexity of the information ecosystem. Although information providers may make documents freely available on the web, there are legitimate differences of opinion as to whether libraries and archives can add them to their collections and provide access to those copies under the current provisions of copyright laws in the United States (especially fair use as defined in 17 USC 107 and the exemption for libraries and archives for preservation of copyrighted materials in 17 USC 108). The Library of Congress National Digital Information Infrastructure and Preservation Program (NDIIPP) and the U.S. Copyright Office are sponsoring the Section 108 Study Group to explore these issues in the context of records creators, distributors, libraries and archives, and patrons.

[36] The Illinois State Library contracted with Larry S. Jackson of the University of Illinois at Urbana-Champaign to develop software based on the open-source wget spider under an IMLS grant. See "GSLIS Electronic Archives Project" at http://www.isrl.uiuc.edu/pep/ (checked 18 February 2007). The Internet Archive at http://www.archive.org/ (checked 18 February 2007) has been harvesting virtually the entire web for many years. It has recently begun offering Archive-It as a commercial service which is being used by some state libraries to capture state agency publications on the web. See http://www.archive-it.org/ (checked 18 February 2007). The California Digital Library's Web-at-Risk project, funded under a Library of Congress National Digital Information Infrastructure and Preservation Program grant (http://www.cdlib.org/inside/projects/preservation/webatrisk/, checked 18 February 2007) and the University of Illinois at Urbana-Champaign's ECHO DEPository Project (http://www.ndiipp.uiuc.edu/, checked 18 February 2007) are two other projects focusing on harvesting documents from websites.

[37] "Archival Reference Services for Digital Records: Experience with the Access to Archival Databases (AAD) Tool," Case Study 9.

Arrangement

Libraries and records professionals generally take very different approaches to arrangement. Librarians typically classify materials from many different sources by subject into a single system, bringing related materials together for convenient access. The Dewey Decimal System and the Library of Congress Schedules are two common schemes used to classify documents.

Records managers and archivists typically do not reorganize the records, but leave them in their original order. This practice is based, in part, on the notion that the records creators organized the files to facilitate their retrieval, and observing provenance and original order preserves the evidence of that structure and function.

Information professionals must reconceptualize what it means to organize their collections in the digital era. Is there value in grouping documents together in a virtual environment? Does original order have meaning in a database that was queried using many different sort criteria? Classifying a complex work under one concept was an inherent limitation when working with physical documents, but easily overcome by using metadata in many categories that point to the same virtual document. At least in an institutional context, the real work of arrangement may involve working with records creators to ensure records are properly classified when they are created.

Still, considering that the original purpose of classification was to facilitate access, one must wonder why records creators would spend any time sorting records to speed retrieval when operating systems and databases can do full-text searches to quickly find the desired record. In fact, if a record is misfiled or the user can't remember the correct category, classification may be more hindrance than help. Yet, full-text searches are not always efficient or successful in retrieving the desired information. Traditional approaches to classification may be superseded by macro appraisal and functional analysis to help manage massive quantities of records.

Description

"Having heard someone once define information science as 'librarianship practiced by men,' Mr. Gorman concluded by defining metadata as 'cataloging practiced by ill-informed men.'"[38] Although pithy and humorous, the reality is that traditional approaches to cataloging are limited at best and irrelevant at worst. Stewart Weibel developed the Dublin Core, in part, as a simplified approach to description that did not require professional catalogers, and the Joint Steering Committee for Revision of *Anglo-American Cataloging Rules* is abandoning that title for

[38] "The Future of Anglo-American Cataloguing Rules," Art Libraries Society of North America 27th Annual Conference, Vancouver, BC, 27 March 1999. Sherman Clarke, panel moderator. Available online at http://www.arlisna.org/news/conferences/1999/proceedings/aacr2.htm (checked 18 February 2007). Cited in *A Glossary of Archival and Records Terminology.*

a new work intended to address the challenges of the digital era, *RDA: Resource Description and Access*.[39]

Description has two goals, physical and intellectual control. Information professionals need an inventory of their holdings so that they can manage the materials. At the same time, patrons need tools to help them find information relevant to their research. The question is, "How can this be accomplished in the digital era?" If there is a fundamental skill for description that information professionals need in the digital era, it is the ability to invent new ways of fulfilling these two goals.

"Description" often connotes formal bibliographic processes used in the analog environment. In the digital environment, its parallel is "metadata." The participants recognized that the two are complementary but not synonymous. Description generally entails the creation of a "surrogate" for the records, and the nature of that surrogate has evolved dramatically. Today description refers to a systematic approach to documenting holdings with an eye to access, and includes descriptive standards, such as *Anglo-American Cataloging Rules* (AACR) or *Describing Archives: A Content Standard* (DACS);[40] data architecture standards for storing and exchanging the descriptions, such as MARC; and data value standards for controlled vocabularies, such as the Library of Congress Subject Headings. These various standards serve to structure metadata about collections.

Archivists adopted EAD and DACS in the relatively recent past to support automated finding aids. Some participants saw these standards as closely tied to an intellectual model of records based on collections of paper records, although many participants at the colloquium expressed their belief that the ability to create EAD finding aids for digital records would remain an important skill.

However, archivists can now rely on other types of description to maintain control over and provide access to their holdings. For example, archivists are frequently faced with creator-generated metadata received with the records. This data will vary from series to series, and archivists must understand how to integrate it with the metadata added at the repository so that the patrons can use those descriptions to identify and select materials to review. Automating the acquisition process can also generate a database that provides physical control for every item transferred to the repository, noting when it was received, the source of acquisition, and any number of other details. In addition, computers can perform full-text searching of documents, and relevancy ranking algorithms have proven that keyword retrieval is often very effective. The case study presented by Elizabeth Yakel and Polly Reynolds described the novel use of technology to "explode the finding aid," to completely rethink the assumptions in DACS

[39] See "RDA: Resource Description and Access" online at http://www.collectionscanada.ca/jsc/rda.html#drafts (checked 20 February 2007).

[40] Society of American Archivists, 2004.

and EAD. Their case study also incorporated concepts of Web 2.0, encouraging patrons using the finding aids to participate in the descriptive process by contributing information.

It is clear that information professionals need to know how to use computers to improve the descriptive process. That means more than technical skills; to use computers effectively, archivists must have the theoretical knowledge of the principles of description and access to ensure that they develop systems that achieve those ends. At the same time, they also need to know enough about technology to find innovative ways to use technology to achieve the goals without merely recreating manual processes.

Given the power of automating retrieval, the real job of description may be providing patrons with information not inherent in the records: context. Information professionals, especially records professionals, may spend little time describing records. Rather, they may be writing administrative histories and scope notes so that patrons can understand what they are looking at and narrow their searches to more appropriate collections or series.

Description in a digital world will require a combination of traditional approaches and innovative new applications. Application of technology is not the only viable approach. Not all documents lend themselves to full-text searching. Neither images nor nonvocal music have text to index. Providing access has always entailed both natural language description and controlled access. The precision of text retrieval often improves when a cataloger assigns controlled vocabulary for concepts that might not be inherent in the text of a document. While information professionals may now think of metadata rather than cataloging, reliance on Dublin Core is insufficient. Dublin Core is a very limited set of metadata elements intended for discovery. If the Open Archives Information System has taught the profession anything, the long-term preservation of digital documents requires a vastly richer set of metadata elements, and those metadata elements address much more than discovery.

Housing: Containers and Storage

Information professionals are faced with two fundamental questions. First, they must determine what kind of container to use to keep digital records, in which format to encapsulate the bitstream. Second, they must determine how and where to store those containers. Information professionals face a rough parallel with paper; what kind of folders and boxes for containers? and what kind of shelving and environment to store the folders and boxes?

Choosing the right container may have significant impact on the content and authenticity of the digital information. Scanning a color document in black-and-white may save storage space, but there's the potential for loss of meaning indicated by color. Converting a spreadsheet to a PDF may capture the readily apparent content, but underlying formulas will be lost. Choosing the right container to store digital records is also closely tied to preservation. The challenge is to find a format that can be read and rendered long into the future, long after the software used to create the record has become obsolete. Information professionals must be able to assess the long-term implications of selecting one format over another.

Participants made frequent reference to XML as the current standard of choice for a container. A few participants reminded the group that for all its benefits, XML would—like all formats—ultimately become obsolete, and that information professionals must try to think beyond the horizon. Participants also mentioned PDF-A. They discussed the importance of understanding both images formats, as well as standards for digitizing documents to ensure that the resulting images were of sufficient quality.

Choice of format, like all tasks, requires that skill be informed by the creative application of knowledge. Digital records may not be stored in only one format; the record might be stored in its original, native format to help ensure the ability to demonstrate authenticity, as well as in a plain-text rendering to support full-text indexing and as a last-resort preservation version. Ultimately, the choice of format is a calculated risk, and it's impossible to predict the future. The more one understands the ultimate goals of storage and the more techniques one is familiar with, the more options to evaluate as the best.

Participants noted that storing digital records raises a fundamental question that has been debated in the professional literature, "Will the repository take custody of electronic records, or will the records remain with the creator in a post-custodial environment?" Participants felt that the answer would differ, depending on the nature of the repository and the nature of the specific record series. However, the answer presumes that libraries and archives must be prepared to take custody of some records and publications.

To do this, participants felt that information professionals responsible for storing records would need to know a fair amount about hardware. At first, it might seem that these skills and knowledge could easily be delegated to someone with technical expertise. In fact, participants recognized that technologists often think about data storage very differently from archivists. The very word "archives" demonstrates the difference in the communities. Where technologists use the word to mean offline storage, assuming that "archived" records are of limited value because they're seldom if ever used and will be deleted in a few years, librarians, archivists, and records managers understand that these records are of exceptional importance and must be kept indefinitely.

Information professionals need a comparative understanding of the nature of physical storage media, such as tape, magnetic disks, and optical media. They need to understand the hardware used to read those media. And, they need to know how those systems are attached to each other and to the network. Ultimately, they need an overarching understanding of the architecture of the storage network.

Participants felt it was important for information professionals to be able to evaluate how well different technologies met their needs. How robust is the system? How does it ensure and protect the integrity of the data? How does it address data recovery, but in terms of hardware failure and environmental disaster? What is the maximum capacity of the system and what will be done when that capacity is reached?

Information professionals also need to understand something about the software used to manage these storage systems. Participants mentioned a number of different systems, including document management systems, content management systems, and the use of basic operating system file systems, such as Windows' FAT32 and NTFS, or Unix's and Linux's NFS, xfs, and ext3fs. Participants mentioned a number of open source systems, including DSpace, Fedora, LOCKSS, and Storage Resource Broker systems.

Participants recognized that storage systems can be extraordinarily complex, especially as the size of the system grows beyond a server or two. In small repositories with limited technical support, the information professional must know some basic system maintenance, such as how to back up the system and verify the integrity of the backup. Participants felt that most systems would quickly grow in complexity and require professional systems administration. However, once the system was designed to meet the needs of the repository, the maintenance could be contracted.

In the best scenario, data will be stored in a distributed environment to address disaster recovery. That demands skills for managing a more complex system, as well as providing security. Participants pointed to the importance of a trustworthy information system[41] that allows the repositories to demonstrate the policies and procedures used to protect the information from corruption (either through degradation or attack) and create audit trails that ensure those policies and procedures were followed.

PRESERVATION

Digital preservation has received significant attention because of the ephemeral nature of digital materials and rapid obsolescence of information technologies. Preserving digital materials poses problems similar to the challenges of preserving traditional media, including security, environment, and disaster planning. However, the problems of preserving digital media are more immediate and pressing. One might find a grandparent's letters and photographs in an attic and still be able to read them; but will future generations be able to read e-mail and digital snapshots on disks, thumb drives, and memory chips?

More often than not, information professionals will help develop policies and procedures to implement preservation measures and then verify compliance with those policies and procedures. In smaller shops, information professionals may be responsible for preservation work, including making and verifying backups, managing media, and changing passwords. In all shops, information professionals may find that they need very practical IT skills to work with obsolete formats, and finding contractors familiar with obsolete technology may be difficult at best.

---

[41] *Trustworthy Information Systems Handbook* (St. Paul: Minnesota Historical Society, 2002), online at http://www.mnhs.org/preserve/records/tis/docs (checked 17 November 2007).

Authenticity

Given the ease with which digital documents can be altered without any readily apparent indication of change—through malicious or unintentional act, or through deterioration—information professionals must pay particular attention to protecting the integrity of their digital holdings.

Archivists, who are often asked to certify records, are particularly concerned about the ability to demonstrate the authenticity and integrity of records. To do that, information professionals must be able to verify the origin of materials transferred to their care, and they must be able to determine if the materials were altered during transmission. They must also be able to demonstrate that the materials have not been altered while in their custody. Participants felt that information professionals should be familiar with hashing technology and digital signatures, two techniques commonly used to ensure that a record has not changed. They must also know how to create or specify audit tools and processes to support validation of the materials.

Protecting Collections

Digital records have storage requirements similar to traditional media; the equipment and media need a climate controlled environment. Digital records may be less susceptible to vermin, they are also subject to online viruses. Introducing a single, contaminated record may result in other records being infected. Virus scanning is a critical skill for protecting digital records.

Information professionals must also ensure that the system is secure. They must assess whether only authorized individuals have appropriate physical and network access to the system.

All information professionals need to be able to develop and maintain a disaster plan that incorporates the repository's technical infrastructure and digital collections. At a minimum, that means having effective, tested backup procedures that include storing a redundant copy offsite. A fully developed plan will also include procedures for business continuity and recovery that will allow the repository to bring its digital collections back online as rapidly as possible and with little or no loss of information.

Keeping Documents Alive

Repositories must find ways to refresh the digital signal over time to prevent bit loss. They must be able to read obsolete formats, such as 8, 5¼, and 3½ inch floppies in their collections. The must know how to address the challenges of different software packages that encode data in different formats; Word may not be able to read WordStar or XyWrite. Software formats, even for the same application, change over time; Word 2003, for example, is radically different from Word 2007.

Participants identified a number of skills to keep documents alive, including media refreshing to keep signals from fading, as well as media and software format migration to counter obsolescent technologies. A number of participants also mentioned emulation as a preservation

strategy. Keeping documents alive effectively, especially when migrating from one software technology to another, may require knowledge of data manipulation and transformation.

Participants mentioned a number of standards that information professionals should be familiar with, including PREMIS[42] and METS.[43] The participants also referred to JHOVE, a tool to identify the binary type of digital objects, such as the creating application and version of that format.[44] PRONOM[45] is a similar project, as is the Global Digital Format Registry.[46]

As software and hardware change over time, the manner in which records are rendered also changes. A document created on one computer may look different when displayed on another computer. In many cases the differences may be trivial; for example, the meaning of a document is no different if rendered in Times New Roman or Arial. However, changes in formatting (bulleted lists, color, or margins) may significantly diminish the usefulness of a record. Presentation is not a new issue. Information professionals have accepted microfilm as an invaluable preservation tool, even if it is imperfect. Most microfilming projects use monochromatic film, with the result that color is lost. Often, the film is high contrast, making it hard to read penciled notes in margins. Information professionals need the skills to assess the impact of changes in presentation.

Although participants all believed that the best strategy for preserving digital materials was to participate in the design of electronic recordkeeping systems, they realized that would not always be possible. Given that rare, obsolete formats may survive principally in repositories with historical collections, most information technologists who work with current formats will not be familiar with these materials. In some instances, information professionals will have to engage in digital archaeology when they receive documents in obsolete formats. Knowledge of computer forensics may be a particularly valuable skill. As a result, there may be a demand for a new specialization in the information and conservation professions.

Reference and Access

Working with patrons may be radically different in the digital era. No doubt many patrons will continue to visit libraries and archives to conduct research in person. Others will contact a librarian or archivist for assistance, either through e-mail or an interactive chat service. But many patrons will find and retrieve materials on their own without ever communicating with an information professional. In many ways, this disintermediated access sounds very much like the

---

[42] Preservation Metadata Implementation Strategies. See http://www.loc.gov/standards/premis/ (checked 1 April 2007).

[43] Metadata Encoding and Transmission Standard. See http://www.loc.gov/standards/mets/ (checked 1 April 2007).

[44] JSTOR/Harvard Object Validation Environment. See http://hul.harvard.edu/jhove/ (checked 1 April 2007).

[45] http://www.nationalarchives.gov.uk/pronom/

[46] Maintained by the Harvard University Library. See http://hul.harvard.edu/gdfr/ (checked 29 July 2007).

fantasy patrons have dreamed for years, to ramble through the stacks where they will find exactly what they are looking for without leaving the comfort of home or office.

Understanding reference and access in a digital environment entails approaches from both the users and the information professional's perspectives. Users need to be able to locate information, and they need to be able to understand what they find. Archivists need to understand who their users are as well as how they attempt to access information.

User Needs

The reality is that most patrons need help finding the information they need. Often they need help framing and focusing their questions. They may be looking for information in an unlikely place, when the information can be found more readily elsewhere. Participants generally assumed that patrons would access archives through the web. To do that, information professionals need to know something about developing web pages using markup languages such as HTML, XML, and TEI.[47] That immediately pointed to the need to understand human computer interfaces and the related skills of developing websites that were easy to use and of testing for usability. That usability includes not only design, but also the organization of information on the site.

And, patrons often have difficulty assessing the quality of information. A part of the solution may be to develop truly effective self service tools, to develop interfaces that lead patrons to the right place through a series of reference questions to address anticipated patron needs.

Archival Responsibilities

To provide effective reference service, information professionals need to know something about their patrons, and the ability to conduct user studies is an important skill to help gain that knowledge. They should also understand how to mine information they already have about users, such as access logs on the web server. Getting information about patrons is particularly important when working on the web; repositories may be reaching entirely new audiences with radically different needs than patrons who visit the repository in person. Repositories may discover they need a suite of access tools for different audiences.

At the same time, participants recognized that information professionals need to take advantage of tools to improve search, based on understanding the information seeking behavior of their users and the changing nature of their holdings. Archival resources are no longer limited to physical collections with recognizable hierarchies. Archives provide access to a wide range of resources, including databases, indices, and virtual exhibits. The traditional finding aid is a superb tool for navigating a physical collection, and participants generally felt archivists should be familiar with EAD and other descriptive standards. However, patrons have found the tradi-

[47] Text Encoding Initiative. The TEI consortium has developed guidelines for markup of linguistic and literary texts. See http://www.tei-c.org/ (checked 1 April 2007).

tional archival hierarchy frustrating and hard to understand without professional mediation. And a single approach will not provide effective access to all holdings.

Researchers have become accustomed to Google, which brings up the information they need, often on the first attempt. Participants recognized that using Google to search the Internet for a quick answer is fundamentally different from the research done in libraries, but they also acknowledge that Google sets patrons' expectations for search. As a result, many participants felt it was important for information professionals to learn about search engines so that they could apply those tools to their own collections. They also need to learn about how commercial search engines rank results so that repositories can expose their collections using techniques to improve how their materials are ranked.

Information professionals must be familiar with emerging trends in an Amazoogle[48] world to transform access to materials. They need to discover and learn how to use new technologies and take advantage of the philosophy of Web 2.0. Beth Yakel and Polly Reynolds' case study looked at the use of referral software and patron participation to enhance an archival finding aid.[49] While the notion of tapping a world of volunteers is attractive, participants raised concerns about the reliability of information provided by unknown contributors on the web. Archives have a reputation for authentic, reliable information, and allowing individuals to post information that has not been checked could damage that reputation when users try to "correct" the record. As stated earlier, patrons have difficulty assessing the quality of information, and including data beyond that created or vetted by the archives raises other issues.

Ultimately, until everything is online (if ever), patrons will likely need to be reminded that "it's all on the Internet" is a myth. Information professionals will be prioritizing materials for digitization for the foreseeable future, leaving an online environment which includes an assemblage of digitized holdings, born-digital records, and digital description. Readers will demand fast and transparent access to all of these resources.

Limits on Access

Participants raised a number of concerns about making their collections widely available on the Internet. Copyright and rights managements were obvious restraints on which materials repositories can post online. At the same time, many records contain confidential or sensitive information that should be protected. Repositories must also take care to protect the identity of patrons using the collections online.

[48] Coined by Lorcan Dempsey of OCLC.

[49] "The Next Generation Finding Aid: The Polar Bear Expedition Digital Collections: A Case Study in Reference and Access to Digital Materials," Case Study 8.

SOFT SKILLS

Although the colloquium was intended to identify technical skills that information professionals need in the digital era, participants kept returning to other, "soft" skills that were also required. More often than not, information professionals needed these skills when working with traditional materials. However, those skills often took on increased importance in a digital environment or had to be used in a different context.

PERSONAL SKILLS AND ATTITUDES

Often, success in working with digital materials seemed less related to a particular skill set. Rather, the situation demanded a certain attitude. It should not be surprising that many participants felt that creativity and a willingness to take risks were more important than knowledge or skills. The profession is on the edge of an unknown frontier, not a place for the faint of heart. As Elizabeth Adkins observed, "People do not want to deal with the unknown because there is no road map."[50]

Thinking Outside the Box

Because work with digital materials demands innovation, information professionals cannot continue to work the same way they have always worked. Professional work demands thoughtful decisions be made to guide the particular circumstances of the work at hand. However, changes in the work processes demand "meta thinking." Participants identified a number of thinking styles that will help information professionals succeed in the digital era.

Abstract and conceptual thinking suggest the way to simplify a problem is to move back from the specific, in order to see the forest for the trees. Information professionals should be able to discover parallels; can best practices tied to tangible records be translated to the virtual world? Can approaches to manage one type of record be applied to other types of records?

Analytical and systematic thinking point to the ability to break down a problem into specific steps and sequences, to understand how those steps relate to one another. Thinking outside the box demands intuition and judgment. It will often be impossible to know which approach is the best. It's impossible to predict the future, to know the "right" path to take. But to the extent possible, information professionals need a sixth sense to guide them.

Attitudes

Beyond specific skills for understanding a process, the right attitude can help information professionals in the digital era. How they view the challenges of working with digital information is critical. The rapid changes in technology and best practices demand flexibility and adaptabil-

[50] Recorder's notes.

ity. Given the current and foreseeable ambiguity of the information ecosystem, information professionals need to find the balance between continued investigation and decisive action; they cannot let the perfect be the enemy of the possible.

Participants felt that information professionals must recognize the limits of their knowledge. They must be committed to continuing education to keep apace.

Creativity and Problem Solving

Rather than seeing a threat, information professionals must see opportunity. Rather than being frustrated by the unknown, they should sense a chance to be creative. Information professionals must be committed to innovation, to find new and better solutions to the complex problems of digital materials. In particular, that creativity needs to be directed to achieving practical results, including adapting existing solutions and discovering novel approaches to new situations.

COMMUNICATION SKILLS

Throughout the colloquium, participants expressed the belief that information professionals could not solve the problems of digital information alone. Many professions have a piece of the puzzle, including information technologists, records creators, and others in the information ecosystem. It will take all those professions working together to solve these problems.

Communication is an essential part of working with other groups. Effective communication is by no means a new skill. Information professionals have always needed to be able to communicate across cultural and professional boundaries. If there are any differences in the digital era, it is that they must be able to talk to even more groups and they must be able to talk in terms that technologists understand.

Advocacy and Outreach

One constant theme throughout the discussion was the importance of ensuring records management and archives functions were part of the specification of electronic information and recordkeeping systems. To achieve this goal, participants knew advocacy was essential. Information professionals will not be invited to participate in the design process unless those designing the system understand why information professionals have a stake in the process and the value of records and have something to contribute.

Information professionals need the skills to develop effective outreach programs that target the right individuals: skills that communicate why they need to be part of this process, and are timed to ensure they are at the table at the beginning of the process. The message must be expressed in terms that motivate the other stakeholders. Historical value may motivate information professionals to preserve information, but an effective cost-benefit analysis will be much more meaningful to those in the business and IT side of the organization.

Collaboration

Given the number of groups in the information ecosystem and the interdependent nature of their work, information professionals can take a leadership role if they have strong skills in building collaborative networks across a wide range of disciplines. The participants recognized the need for records managers, archivists, and librarians to work together, but they also recognized that those professionals needed to work more closely with disciplines far afield. They may need to take the role of translator, helping different groups with different motivations and values understand each other. Information professionals may benefit from serving as conveners, facilitators, and team builders.

Team Building

Participants identified team building as a specific skill that information professionals need. They also noted that information professionals must be willing to "share turf," to acknowledge that, in some instances, others may be in a better position to do something traditionally identified with their own discipline.

Relationships

In many ways, effective communication and collaboration is more than sharing information. It means building good relationships that transcend the task at hand. It is unrealistic to expect that all collaborators will become friends, but information professionals can strive to build positive, supportive work relationships.

Unless an individual knows another person's interests, it is unlikely he or she will think to include the other when working on a project of interest. If information professionals want to be included in discussions of technology projects, they need to find a way to let those working on technology projects know of their interests. Often those individuals are not in technology, but are responsible for specific programs. By the time IT is brought in to implement a project, it may be too late to address recordkeeping issues.

Building relationships means more than a simple introduction. It's more than understanding information professionals' interests. It's a personal relationship. Individuals may not think of recordkeeping or the library, but they may think of the record manager or the librarian.

Managing Change

Finally, participants recognized that if information professionals want to be successful leaders in such rapidly changing times, they must help those around them cope with the new and unfamiliar. They must help manage change.

REFLECTIONS

The intent of the colloquium conveners was to build consensus on a specific set of skills that an information professional would need to be successful in a digital environment. The results of the colloquium confirmed that archivists are certainly not a monolithic group. Participants posited different skills on the basis of a widely different set of environments. Archives come in so many sizes and flavors, and consistency does not even exist within a specific category of repository, such as college and university archives. Individual efforts tend to lack coordination. Thus, instead of presenting a checklist of knowledge and skills, we offer a discussion of the range of topics and issues that information professionals will address in their work, although the extent and level will vary greatly among institutions. In addition, it became clear that the crucial areas included both technical (hard) skills and facilitative (soft) skills. One dramatic consequence of technology is the removal of the walls of the archives. No longer do archivists operate in isolation.

We must recognize that moving forward entails acceptance of our differences. The lessons about electronic archives drawn from government repositories will not necessarily apply to smaller, private repositories. On a state government level, for example, archivists work with selected key partners to develop standards and guidelines that form parts of the state's technology infrastructure, emphasizing statute and policy. This work will not enable the archivist in a small manuscripts repository to deal with the personal papers of a famous poet.

Consensus among the participants focused more on the needs we share than the specific skills we must have. These commonalities should shape the next round of discussions.

Evaluation

Cleary we need more investigation and analysis. The case studies represent projects in various stages of completion. Thus, presenters could not always offer sets of objective data and conclusions. Completed projects with documented results will help others adapt approaches to different situations. Completed projects can also highlight the skills whose presence or absence affected the success of a given project.

Communication Skills

Throughout the colloquium participants identified communication as essential to working within any environment. In a larger sense, communication is crucial if we are to learn from what our colleagues are doing. The colloquium was a prime example of an opportunity for exposure to a range of projects, perspectives, and opinions, and for the case study presenters to gain perspective from the comments of others.

Collaboration

Communication should lead to collaboration. As boundaries become more blurred, collaboration has become another valued skill within institutions. On a macro level, participants realized that we benefit from collaborative initiatives that can take advantage of the skills of others. Within each case study, the need for collaboration was clear.

The New Skills colloquium may not have produced a definitive set of skills. In hindsight, the colloquium suggests no single skill set will fit all jobs, although it identified a myriad of issues information professionals face in their work in an electronic environment. We have moved the discussion forward and identified the kinds of questions information professionals must ask. Now we must begin to answer those questions by collaborating within our organizations and across them, evaluating the successes and failures of that work, and communicating the results of those case studies on a regular basis. Soon, sufficient commonalities will emerge that will enable identification of best practices and standards that represent the best thought of the information professions, and real world solutions for our constituencies.

APPENDIX 1
KEYNOTE ADDRESS

Are We Ready for New Skills Yet?

Margaret Hedstrom
University of Michigan

It can be useful to frame the question of what new skills archivists need in the digital environment around what we know about the nature of skill from the perspective of sociology, cognitive psychology, organizational studies, and economics. The nature of skill and its relationship to organizational capabilities, such as memory and learning, has been the topic of considerable research in the social sciences during the last three decades. One could attribute the scholarly interest in skill to fundamental changes in organizations and the economy as we shift from the industrial age to the information age. To set a larger framework for the question of what new skills archivists, librarians, and records professionals need in the digital age, it is useful to consider the questions What is skill?, Why are skills important?, How does one recognize skill?, and How are skills taught and learned?

Richard Nelson and Sidney Winter devoted an entire chapter of their now classic book, *An Evolutionary Theory of Economic Change*, to the topic of skills.[51] They defined *skill* as "a capability for a smooth sequence of coordinated behavior that is ordinarily effective relative to its objects, given the context in which it normally occurs." According to their definition, skill is not limited to the efficient performance of repeated and practiced tasks. Skill also includes "choice behavior" where analysts work with imperfect and fragmented information, deliberate, contextualize, and sometimes rethink organizational goals.[52] Nelson and Winter also identified three characteristics that many skills have in common. First, skills are programmatic because they follow a sequence of steps where each new step is triggered by completion of the previous step. Second, skill is built on tacit knowledge where a skilled performer may not be fully aware of the details of the performance and may have difficulty articulating exactly what constitutes his or her repertoire of skills. Finally, skilled performance entails the exercise of many choices, but the options are "preselected" and choices are made with little awareness or contemplation.

The "tacit-ness" of skill and the embedded nature of choice in both individual behavior and organizational routines make articulating, teaching, learning and legitimating skills challenging. The skilled trades and the professions often turn to external signifiers, such as professional de-

---

[51] Richard R. Nelson and Sidney G. Winter, *An Evolutionary Theory of Economic Change* (Cambridge, Mass.: Harvard University Press, 1982), chapter 4, 73-95.

[52] Nelson and Winter draw on earlier work by Herbert Simon, Richard Cyert, and James G. March on the concept of bounded rationality rather than optimization in decision making. See Herbert Simon, *The Sciences of the Artificial,* 3rd ed. (Cambridge, Mass.: MIT Press, 1996); and Richard Cyert and James G. March, *A Behavioral Theory of the Firm,* 2nd ed. (Cambridge, Mass.: Blackwell Business, 1992).

grees, certification, accreditation, and licensing to demonstrate competence and to police the boundaries between their areas of jurisdiction and competition from both unqualified practitioners and other trades or professions that might lay claim to their area of practice.[53] Librarians, archivists, and records managers have used these tactics, not only to ward off non-professionals, but also to prevent other professions, such as systems analysts, lawyers, and information managers, from encroaching on *our* territory. These battles have not been limited to so-called outsiders. Librarians, archivists, and records managers have debated among themselves over who owns which part of the life cycle, which types of material, and which audiences.

People who study the nature of skill also analyze how skill is acquired and transferred. Carl Polanyi observed that one difficulty in acquiring skill is that "the aim of a skillful performance is achieved by the observance of a set of rules which are not known to the person following them."[54] People typically learn new skills through a process of observation and then a disciplined regimen of repeated practice guided by an expert instructor. Any of us who have learned (or failed to learn) how to play a musical instrument, excel in a sport, or speak a foreign language fluently can attest to the importance of guided instruction and practice. As Nelson and Winter point out, verbal instruction, embodied in the how-to-do-it manual, is only the starting point for skill acquisition because the manual includes only the articulable portion of the knowledge involved; it does not certify possession of the skill.[55] This is why most of us would be reluctant to fly on an airplane piloted by an individual who had read the instruction manual on how to fly an airplane but had never practiced; or go to a surgeon who had memorized the steps in a surgical procedure but not performed any surgeries. Learning and transferring skill is a challenging problem that typically is addressed by attempting to make tacit knowledge explicit (as is the case in knowledge management and expert systems) or by circulating skilled practitioners so that others can observe and absorb their tacit knowledge.[56]

Skill in knowledge work is doubly difficult to observe and attain because the skilled knowledge worker performs smoothly and efficiently by making as many choices as possible "automatically" and by knowing how to rule out choices and place upper bounds on the number of options to consider. In many types of knowledge work, the available repertoire of actions and decisions is constrained by rules, codes, procedures and best practices, but even then professionals choose the most appropriate actions based on their knowledge and past practice. Skilled practitioners know how to limit their decisions to what is practical, realistic, and likely to produce a satisfactory (although not necessarily optimal) outcome. For example, the choices that an architect makes in designing a building are constrained by the law of gravity, the local building codes, good architectural practice, and the budget of the client. This allows architects

---

[53]Andrew Abbott, *The System of Professions* (Chicago: University of Chicago Press, 1988), 59-85.

[54] Polanyi, 1962, quoted in Nelson and Winter, 77.

[55] Nelson and Winter, 77.

[56] Linda Argote, *Organizational Learning: Creating, Retaining and Transferring Knowledge* (Boston: Kluwer Academic Publishers, 1999), 87-88.

to focus their attention on matters of design, functionality, aesthetics, and the suitability of the proposal to the client's needs, desires, and money.

Because skills are tacit and, like infrastructure, are most obvious when they break down, disruptive change has a way of exposing what the traditional skills were, but disruptive change does not predict very well what new skills are needed to replace those that have become obsolete or which traditional skills have lasting relevance to professional practice. We are witnessing a period of unprecedented change in the nature of individual and organizational communications, work processes, and decision-making. While many of these changes are linked in some way to the use of information and communication technologies, their implications for archival theory and practice are more complex. As a consequence, there are many new areas of professional practice that lack a useful theoretical basis or replicable practices and skill sets. These disruptive changes should trigger a reassessment of the skill set of archivists, records managers, and librarians.

Let me turn now to the title of this talk: Are We Ready for New Skills Yet? Or do we have a sufficient knowledge base to build skills? My answer to these questions is decisively yes and no. I'll remind you of Nelson and Winter's definition of skill as "a capability for a smooth sequence of coordinated behavior that is ordinarily effective relative to its objects, given the context in which it normally occurs," and reiterate their three characteristics of skill as programmatic, built on tacit knowledge, and decision intensive. I think it is safe to say with regard to our responsibilities for digital information that archivists, records managers, and librarians lack both the formal and tacit knowledge necessary for refining and developing all of the new skills we need. In fact, I would argue that given the disruptive changes in technology, organizations, and communication, it might be premature to commit to a hardened skill set at this point in our evolution. This is not to say that we have no idea of what skills we need or that we can't begin to build up a skill set, but this framework suggests that we should look at skills as dynamic and evolving (at least for now).

To develop skill in a "programmatic" sense it is necessary to have a clear sequence of steps that one can practice and learn. Often, when dealing with digital information, archivists, records managers, and librarians encounter novel problems that they have never encountered before. These problems may include new and unfamiliar technologies, new terminology and concepts, new communities with which to interact, and new ways of doing work. Problem solving itself then becomes an important skill out of which a new set of practices may emerge. To move beyond the problem-solving stage to a fundamental shift in professional knowledge, practice, and skills, it will be important for those professionals involved in problem-solving to share their approaches, results, successes, and failures. Each attempt to solve a problem needs a well-designed evaluation that can isolate success factors and points of failure. Building a conference like this one around case studies is a good starting point, but more systematic reporting on how different organizations have addressed similar problems would start to build a foundation for systemic change. Problem-solving alone can turn into a huge exercise in trial and error if we don't also conduct research into a wide range of issues and develop foundational knowledge on which to build new practices and skills.

To make this more concrete, I will suggest a few examples of areas where research is essential. Understanding the changing nature of contemporary organizations is one critical area for research. Much Western archival theory is based on a model of organizations from the last century that conforms to a strict hierarchy, a rigid separation of responsibilities and functions, a clear division of labor, and predictable and finite flows of information through the organization and between one organization and another.[57] Many contemporary organizations deviate considerably from this model. Organizations have been characterized as flatter, less centralized, and more flexible today than they were in the past. Teamwork, project-based initiatives, and collaboration are replacing formal administrative procedures as common ways of organizing work. While bureaucratic routines and their associated paperwork have not disappeared, new forms of coordination and production have radically altered workflows, communications, and organizational documentation. Administrative tasks such as scheduling, correspondence, taking minutes, distributing documents, and managing files increasingly are integrated into the workflows of professional and technical staff.

Equally important is a need to analyze how various forms of electronic records are generated, organized, valued, managed, and kept—often without centralized direction or control. Through a combination of reductions in administrative and support staff and the introduction of electronic recordkeeping systems, records management, to the extent that it exists today, is either deeply embedded in discrete business processes or left to the discretion of end users. Efforts by records managers and archivists to introduce centralized records management programs in contemporary organizations or to reinvigorate dormant ones have met with little success. Rather than attempting to impose a programmatic approach that is out of step with the ways that contemporary organizations actually work, we need research about recordkeeping behavior in organizations in order to design interventions that contemporary organizations can implement and that offer measurable benefits which are worth the investment of effort necessary to make them work.

Another critical area for research concerns the shift from physical to online access to archival materials and the attendant changes in use patterns and user behavior. Prior to any massive conversion of physical archives, we are likely to see an increasing portion of descriptive data and finding aids made available for online searching along with a steady increase of born-digital content online. Use patterns and user behavior are prime areas for research because we lack sufficient understanding of how users discover archival materials, select which materials they would like to use, and interpret what they find. Certainly, some of the literature from related fields is important and helpful, but archivists have little empirical basis for making design decisions about archival access systems based on unique archival requirements. This is especially problematic at a time when some portion of access is shifting from a process that was heavily mediated by reference staff in an archival institution to one where users begin their search for archival materials remotely without benefit of an archivist to assist them. Developing

[57] This type of organizational structure and its recordkeeping practices is discussed most thoroughly in JoAnne Yates, *Control Through Communication*.

a deeper understanding of use and users is also important when archival institutions are making significant investments in revising finding aids and building online access systems.

Digital preservation is the final area that I will mention, although there are many more questions that warrant research. Two recent reports sketch out rich research agendas in this area stressing the need for extensive multidisciplinary research that spans archival and information science, computer science, organizational studies, public policy, and economics.[58] Both reports provide insights into the need for research on digital preservation that is driven by practical concerns about accountability in the public and private sector, the ability to conduct scientific and historical research over extended time frames, and the importance of digital preservation for cultural heritage and personal satisfaction. Although there is ample room for research on technological issues, the reports also stress the need for modeling and redesign of curatorial processes, the development of formal models and metadata schema for representing digital objects and collections, and research on policy and economic models. Digital preservation is gaining recognition as a critical issue for the sustainability of digital libraries and the development of scientific data repositories. The need for an archival perspective on this problem is especially acute at a time when funding agencies are beginning to invest in research on this problem.

In closing, I offer a mini-case study based on my experience with the NHPRC-funded project on Documenting the History of Internet 2—a project that we conducted in collaboration with the Charles Babbage Institute on the History of Information Processing at the University of Minnesota (which ironically has no electronic records). We learned quickly that the idea of conducting an inventory of Internet 2 records, dividing them into record series, and noting their content and quantity was an unfeasible undertaking, even though the project has a half-time graduate student research assistant. Why? Because knowledge of the records was tacit, fragmented, decentralized and incomplete. The two "repositories" set up for centralized recordkeeping were rarely used and the staff of Internet 2 consists of the core administrative staff in Ann Arbor, a smaller DC office, and faculty who work on contract (often part-time) at dozens of universities. Researchers and volunteers do much of the conceptual work and they make extensive use of list serves and collaboratively authored papers. So, we rapidly redesigned our approach to the involved key functions, accomplishments, and events (like the annual meeting) to form the framework for our analysis. Gathering this information involved skills in interviewing people, following leads, and (of course) familiarity with the various technologies and systems.

We also learned that appraisal and selection at high degrees of granularity is expensive and quite possibly unnecessary. Instead of developing a fine-grained set of appraisal recommendations, we gained permission to crawl the I2 website. In this process we learned how to use the free open-source Heritrix crawler available from the Internet Archive and we learned how the

---

[58] *It's About Time: Research Challenges in Digital Archiving and Long-term Preservation* (Washington, D.C.: Library of Congress, August 2003). Available online at http://www.digitalpreservation.gov/index.php?nav=3&subnav=11; and *Invest to Save: Report and Recommendations of the NSF-DELOS Working Group on Digital Archiving and Preservation* (2003). Available online at http://delosnoe.iei.pi.cnr.it/activities/internationalforum/Joint-WGs/digitalarchiving/Digitalarchiving.pdf.

results vary depending on the starting nodes and parameters of the crawl. We experimented with various ways to search and view the data we collected using readily available off-the-shelf hardware and software. To begin to generalize this experience, it would be useful to compare our findings with other organizations that have used similar approaches and technologies. And, I might add, there were many issues we were unable to address: what to do with and about e-mail, how to advise I2 on retention and disposition (because we could not find a process that fit with their work culture) or how to sustain the activity after the grant ended (in part because of staff turnover).

In conclusion, it might be helpful to think of the development of new skills as a dynamic process with high degrees of interaction between research, problem solving, generation and testing of new knowledge, application and evaluation, articulation of formal knowledge, and finally development of tacit knowledge through practice. That last element—developing tacit knowledge through practice—is where we have the longest way to go. Good luck with your workshop and—whatever you do—practice what you learn here when you return home.

APPENDIX 2
CASE STUDIES

ACQUISITION (SELECTION AND SURVEYS, TRANSFER AND INGEST)

1. *Changes in Acquisition: A Guide to the Ingest of Electronic Records*
   Eliot Wilczek, Tufts University
   Kevin Glick, Yale University

   This case study focuses on the function of Acquisition, defining it broadly to include appraisal and accessioning activities. The study works under a broad notion of technical skills to include what resources an archives must have as an administrative unit to undertake a trustworthy acquisition (ingest) process. These resources include policies along with hardware and software in addition to the theoretical and technical knowledge of its staff.

   The case study is based on an Ingest Guide developed by Yale University and Tufts University as part of its NHPRC electronic records research grant project, Fedora and the Preservation of University Electronic Records (2004-083). A draft of the Guide is available at: http://dca.tufts.edu/features/nhprc/reports/3_1_draftpublic3.pdf.

The Ingest Guide refers to ingest broadly, defining it as the entire process of moving records from a recordkeeping system to a preservation system. In this process, the Producer and Archive define the scope of records to be ingested. Then, they agree to the manner of the transfer, validation, and transformation. Only at this point will the two parties be ready to proceed with the actual activities. Each part of the Ingest Guide includes a narrative summary, a flowchart illustrating all of its steps, a description of each step, and the components, resources, products, and documentation necessary to undertake each step.

In using the Guide to undertake acquisitions, archivists will need skills that will enable them to undertake the ingest steps, as well as to create or modify the resources that support those steps.

The case study examines two acquisition scenarios:
Scenario One: A hypothetical scenario of a university archives that undertakes a recurring set of routine, semi-automated accessions according to the steps of the Ingest Guide.
Scenario Two: A real-life scenario of digital video records transferred to a university archives on a series of DVDs according to the steps of the Ingest Guide.

The case study explores the skills needed for each scenario by outlining in a table the parts of the Ingest Guide and identifying the skills needed to undertake the steps and/or develop and use the necessary resources. This description of skills is not part of the Ingest Guide and has been prepared solely for this colloquium.

Skills are grouped into two categories:

Technical

Includes knowledge of file formats, recordkeeping systems, programming languages; systems, network, database administration; schema creation capabilities; familiarity with validation tools.

Archival

Includes theoretical and practical knowledge of archival appraisal, processing, description, and ethics; understanding of copyright law; use of metadata; and management and administrative capabilities.

Scenario One

After every meeting of the University Board of Trustees, the Office of the Board of Trustees (the Producer) transfers a CD containing the working files of the full board and committee meetings to the Digital Collections and Archives (the Archive). These records include agendas, minutes of the previous meetings, various reports, and a video message from the University President. Trustees receive these documents from a secure website a few weeks before each meeting. The records on the CD are in MS Word, PDF, and QuickTime Movie file formats.

The Producer and Archive have agreed to a Submission Agreement that calls for the Producer to transfer a set of working files to the Archive shortly after each Board of Trustees meeting on CDs via hand delivery with the records in a particular type of file format. The Archive describes in the Agreement its planned validation and transformation activities. This is a serial agreement that applies to all successive acquisitions of the working files, allowing both parties to skip Section A of the Ingest Guide, Negotiate Submission Agreement, and focus on the actions described in Section B of the Ingest Guide, Transfer and Validation. Therefore, Section A of the table below focuses on the initial establishment of the Submission Agreement.

Scenario Two

Graduate education in Art and Architecture being particular strengths of University, the Archive actively collects in two different record units. The first record unit, "Material of the School of Architecture," concerning events and exhibitions, consists of posters, programs, brochures, and photographic images (print and digital) documenting events and exhibits. The second record unit, "Art, architecture, and art history theses and projects," consists of research papers, theses, and projects by students. When initially created, the archival materials created and accessioned in these units were almost entirely on paper or three-dimensional objects. Today the Producer submits records upon the completion of an exhibition or the completion of a particular student project. The Producer transfers records to the Archive often.

The Producer and Archive have agreed to an informal Submission Agreement that allows for complete academic freedom, but also leads to a highly variable and often unstructured accession process. The Producer does not necessarily conform to specific terms and conditions of transfer, transferring records without a set schedule, in file formats that may not be feasible for the Archive to preserve, and often not describing the accession sufficiently. This is not a serial agreement and a new Submission Agreement is periodically necessary.

The Producer normally transfers records to the Archive in batches of between 1 and 100 CDs or DVDs. These disks contain a wide range of record types (from flat text reports to three-dimensional models to digital still image and video) and a number of different file formats.

Ingest Guide	Scenario One	Scenario Two
A Negotiate Submission Agreement		
A1 Establish Relationship	• **Archival** skills *(appraisal)* to establish relationship with Producer as appropriate archive for Producer's records and develop and interpret *Records Authority Statement* that serves as warrant for Archive-Producer relationship. • **Archival** skills *(descriptive* and *metadata)* to create a *Producer Record* entry for Producer. • **Archival** skills *(descriptive* and *metadata)* and **Technical** skills *(XML schema creation)* to create machine-readable syntax for *Producer Records*.	
A2 Define Project	• **Archival** skills *(appraisal)* to agree on scope of acquisition with Producer. • **Archival** skills *(appraisal)* to start a preliminary *Survey Report*. • **Archival** skills *(appraisal)* to determine Producer has appropriate authority over records based on *Producer Record* and preliminary *Survey Report*.	
A3 Collection Information and Assess Value of Records	• **Archival** skills *(appraisal)* and **Technical** skills *(evaluation of electronic recordkeeping systems)* to survey records, determining their essential elements, authenticity, appropriateness for accession. • **Archival** skills *(appraisal)* and **Technical** skills *(evaluation of electronic recordkeeping systems)* to create the S*urvey Instrument*, *Recordkeeping Evaluation Tool*, and *Records Retention Policy* to make this assessment. • **Archival** skills (*authenticity requirements)* to analyze and judge the grounds for presuming the authenticity of the records, which is particularly important with these Trustee's records.	• **Archival** skills *(appraisal)* and **Technical** skills *(file formats)* to survey records spread over a number of disks and stored in a number of complex file formats, and collect information about the records necessary to make an appraisal. This particular Producer happens to be on the cutting edge of multimedia technology and the Archive would be required to determine the records' essential elements, authenticity, and appropriateness for accessioning. Also, the Archive must have the skill to understand how to represent the essential elements of form for each arcane format that may need to be converted to place the files in storage or even to view them. Without this step the assessment of the feasibility of preservation cannot be completed.
A4 Assess Record Types	• **Archival** skills *(appraisal)* to determine if the files are not one of Archive's established record types. • **Archival** skills *(descriptive* and *metadata)* to create a *Record Type Record* for any new record type.	

	• **Archival** skills *(descriptive* and *metadata)* and **Technical** skills *(XML schema creation)* to create machine-readable syntax for the *Record Type Records.*	
A5 Assess Formats	• **Technical** skills *(file formats)* to determine if some records are not in formats that meet the Archive's *Formats Standards Policy.* • **Archival** skills *(appraisal)* and **Technical** skills *(file format)* to create the following *Transformation Plan*: Transform Word files to PDFs; keep PDF files as PDFs because that is already a preservation format for the Archive; manage QuickTime files natively, creating a new preservation format. • **Technical** skills *(file format* and *XML schema creation)* to adjust *Format Standards Policy* written in machine-readable syntax.	• **Technical** skills *(file formats)* to determine if some records are not in formats that meet the Archive's *Formats Standards Policy.* In this scenario, the Archive must also be able to reassess as necessary the Archive's *Formats Standards Policy*. Many formats might be considered ephemeral or experimental in nature, but are necessary to preserve the essential elements of some of these records. In such cases, the Archive must be able to weigh the difficultly and expense of one preservation strategy over another (perhaps many of these files should be stored in their native format in order to preserve their look and feel). This would require expertise in determining the digital preservation strategy for that particular format. • **Archival** skills *(appraisal)* and **Technical** skills *(file format)* to create any necessary *Transformation Plan.* • **Technical** skills *(file format* and *XML schema creation)* to adjust *Format Standards Policy* written in machine-readable syntax.
A6 Assess Identifier Rules	• **Archival** skills *(appraisal)* to determine that neither scenario requires the Archive to preserve any *Producer Naming/Identification Scheme* and to determine the appropriate *Archive Naming/Identification Scheme* for the records. • **Archival** skills *(description* and *metadata)* and **Technical** skills *(XML schema creation)* to create *Archive Naming/Identification Scheme*.	
A7 Assess Copyright	• **Archival** skills *(appraisal* and *copyright)* skills to determine that the owns the copyright to the records and based on its *Copyright Policy* it does not need to acquire the copyrights or a license to them. • **Archival** skills *(copyright)* to create *Copyright Policy*.	• **Archival** and **Technical** skills *(digital copyright)* to understand the copyright issues surrounding DVD movies, which are a significant portion of the records in this scenario. Copyright restrictions (from the DMCA for example) may prevent the Archive from fulfilling its *Formats Standards Policy* or *Transformation Plan,* or prevent the Archive from

		"ripping" the DVDs in order to store the records on central storage.
A8 Assess Access Rights	• **Archival** skills *(appraisal, ethics,* and *management)* to determine, based on the Archive's *Access Controls Policy*, that the records should be assigned to a *Record Security Profile* that restricts access to the records for an appropriate period of time. • **Archival** skills (management) and **Technical** skills *(XML schema creation, recordkeeping system,* and *network)* to create this new *Record Security Profile* and determine that the preservation system can accommodate these access control needs. • **Archival** skills *(management and ethics)* to create *Access Controls Policy*.	
A9 Assess Recordkeeping System	• **Archival** skills *(appraisal)* and **Technical** skills *(recordkeeping system)* to use its *Recordkeeping System Evaluation Tool* to evaluate the web environment where the Producer keeps the Trustee working files and the process for copying the working files to CDs.	• **Archival** skills *(appraisal)* and **Technical** skills *(recordkeeping system)* to use its *Recordkeeping System Evaluation Tool* to evaluate the web environment where the Producer keeps the Trustee working files and the process for copying the working files to CDs.
	• **Archival** skills *(appraisal)* and **Technical** skills *(recordkeeping system)* to create a *Recordkeeping System Evaluation Tool*.	
A10 Assess Feasibility	• **Archival** skills *(management)* and **Technical** skills *(systems, network, database administration)* to determine, based on its current *Preservation System Capabilities* report, that this acquisition is feasible.	
A11 Finalize Submission Agreement	• **Archival** skills *(management)* to finalize and endorse the Submission Agreement. • **Archival** skills *(appraisal* and *management)* and **Technical** skills *(XML schema creation, network administration,* and *recordkeeping system)* to create metadata requirements, transfer procedures and schedule, validation procedures, and SIP creation procedures for the records in the acquisition. All of these procedures become part of the Submission Agreement.	
B Transfer and Validation		

B1	Create and Transfer SIPs	• **Archival** skills *(metadata)* for the Producer to follow the Archive's SIP creation procedure, create the SIP(s), and transfer them to the Archive for each recurring acquisition of the Trustee working files. This includes copying all appropriate working files onto a CD, clearly labeling each document according to a specified syntax, clearly labeling the CD as belonging to a particular Trustee's meeting, and hand delivering the CD to the Archive within a few weeks of each Board meeting.	• **Archival** skills *(management)* to act on behalf of the Producer if (as is the case in this scenario) the producer is unable or unwilling to produce SIP(s) according to the terms and conditions of transfer.
		• **Technical** skills *(programming languages, XML schema creation,* and *validation tools)* to create tools and the processes to package the SIP(s). For example, during SIP creation, the Producer or Archive will need tools to create baseline measurements for integrity checks that will be verified later on.	
B2	Validate	• **Archival** skills *(appraisal)* and **Technical** skills *(file formats* and *validation tools)* to check the SIP of each recurring acquisition for viruses; to check the success or failure of the file transfer; to check that the files are well-formed, that the Producer is authorized to transfer the CD, that the SIP contains all of the necessary records components, and that the components all validate. • **Technical** skills *(programming languages, XML schema creation,* and *validation tools)* to create tools and the processes to do this validation work.	
B3	Transform and Attach Metadata	• **Technical** skills *(file format*, *programming languages,* and *XML schema creation)* to transform the records that were slated for transformation in Part A5 (this may require converting files from their native format to a preservation format) and to automatically attach metadata it infers from the Submission Agreement. The Archive would need sufficient skills to maintain file format conversion tools for all of the formats listed in the *File Formats Standards Policy*.	
B4	Formulate AIPs	• **Technical** skills *(XML schema creation,* and *file formats)* to turn the records into AIPs according to its *AIP Configuration Rules*. • **Technical** skills *(XML schema creation, file formats,* and *programming languages)* to create *AIP Configuration Rules* and semi-automated processes to generate AIPs.	
B5	Assess AIPs	• **Archival** skills *(appraisal* and *metadata)* to assess a sample of the AIPs to ensure that they have the records they are supposed to contain.	
B6	Formally Accession	• **Archival** skills *(management)* and **Technical** skills *(system)* to deposit the AIPs into the appropriate preservation system and formally notify the Producer of the acquisition. Specific system administrative and/or programming skills may be required to maintain and/or connect to the preservation system.	

2. *Conducting an Inventory of Electronic Records*

Geof Huth, Government Records Services, New York State Archives
Ann Marie Przybyla, Records Service Development, New York State Archives

This case study examines the complexities of conducting an inventory of electronic records and the questions generated during that process. It is based on findings from a collaborative project—still ongoing—between the New York State Archives and Warren County, a local government in upstate New York. What we are discovering during this project is that many of the underlying assumptions and principles that have guided archivists and records managers may no longer hold true because of the dynamic, integrated nature of electronic recordkeeping systems. We already recognize that we have to rethink our concept of records and records series, and evaluate and adapt such fundamental records management functions as appraisal, scheduling, and destruction. The case study addresses the skills, both technical and otherwise, we in the profession need to develop to accommodate electronic records, and concludes with a series of discussion questions.

Value of inventories to archival organizations

Since 1990 the New York State Archives has provided grants to local governments to initiate and improve their archives and records management programs. As an initial step in that process, local governments across New York State have long used our grants to inventory hardcopy records. In fact, the Archives previously required that an inventory be the first grant project for any local government.

Our reasons for emphasizing records inventories are many. An inventory is the first step in addressing a records backlog, allowing records creators to move inactive records to storage, identify those records they can legally discard, and identify and appropriately manage those records with enduring value. If done effectively, a records inventory yields data that allows an organization to plan for space, identify vital records, design appropriate preservation and conservation measures, and implement finding aids. Records creators who have inventoried their records—and used the results of that inventory to identify records management needs and formulate a plan—are more likely to have records that are pared down to the essentials, in better physical condition, and more intellectually accessible. This is of enormous value to the archives that will eventually accession some of the records and to those archivists who also provide records management support to their customers.

Our strategy for encouraging records inventories appears to have been successful. In a recent survey we found that out of 700 local government respondents, 71% had inventoried their paper records, 76% had developed and were maintaining a records management plan based on a records inventory, and 83% regularly dispose of hardcopy records according to our records schedules. The survey also revealed some troubling statistics: 88% of responding governments did not have a working inventory for their electronic records, 83% did not dispose of electronic records according to state requirements, and 83% did not have policies or a plan that addressed electronic recordkeeping systems. This is of particular concern because the records involved, more than ever, require careful management from the point of creation.

We are now working to close this gap between paper and electronic records inventories. In an effort to understand fully the difficulties of inventorying electronic records, State Archives staff have partnered with a county government to conduct an inventory of the county's electronic records. This project has many goals. First and foremost, it allows us to examine and answer technical questions surrounding the inventory of electronic records. Second, it functions as a staff development opportunity for both state and local government employees. Third, we hope to test our own publications, workshops, and other services to local governments to determine what we must adjust to meet the demands of the digital world. Fourth, we plan to deliver a set of quality products to our county partner, including inventory data, a records management needs assessment, and a records management plan for electronic records.

What we do remains the same, but...

This case both supports and contradicts the notion, "What we do remains the same, but how we do it changes." We are still dealing with records and the appropriate management of information. We are attempting to address yet another backlog of records at the local government level, even though our involvement is more hands on. We rely on the same framework of state and federal law to guide our actions. As records managers and archivists, we still depend on our communication skills, persistence, and ability to collaborate with others. However, there are significant differences in what we are doing, and these differences require new skills, both technical and otherwise, as outlined further below.

New Skills

Relationship skills

The nature of our relationship with our constituency has changed. We are partners—and even students—of our records creators and information technology staff rather than mentors, advisors, and trainers.

Traditionally, State Archives staff have worked primarily with records management officers (RMOs), individuals who, by law, are required to coordinate the records management programs of their governments or agencies. With electronic recordkeeping, however, we must work more closely with consultants, vendors, and other government staff (IT staff, engineers, planners) with whom we've previously had little or infrequent contact. Fostering these relationships—while at the same time encouraging RMOs to focus on information technology issues—is a considerable challenge. We need to make more symbiotic the relationship between IT (which focuses on tools to meet immediate needs) and records management (which focuses on information over the short and long term).

Evaluative skills

We hope to use this project to revise our existing inventory guidelines or to formulate new products to meet the challenges of electronic recordkeeping. We have already noted that we either have to revise or add sub-fields to our existing inventory worksheet. We may need to de-

velop a publication and workshop that specifically address how to conduct an inventory of electronic records. This requires that we evaluate and change what we've always done, a change that can be difficult when dealing with long-term staff and established procedures.

We need to identify ways to improve paper systems that relate to an electronic system. For example, we found that the county clerk's office organized and bound records by series (e.g., deeds, liens, lis pendens). Quadrants of the clerk's "records room" are identified by the types of volumes shelved there, and data entry is performed and tracked by individual series. The electronic system essentially negates this physical arrangement by series because it provides randomized intellectual access. Eventually, the county clerk can file all records numerically as they are received, regardless of the series to which they belong. This electronic system, thus, creates a new set of expectations about access even to the paper records, one that will be difficult to satisfy in a hybrid recordkeeping system.

Technical skills

In addition to the usual arsenal of skills, we have to cultivate our ability to adapt or enlarge our vocabulary, and to accept that we don't always completely comprehend what we're doing in this area. We already foresee a need to acquire real technical skills, maybe not at the level of IT professionals, but more than archivists and records managers in general now have.

We need the ability to understand different operating systems, even though we are dealing almost entirely with Windows. During the inventory process we have found it useful to be able to navigate through file structures to understand how records were created, grouped, and stored, and determine size. In instances when a computing application is not accessible as a Windows application, we need support from the county IT department. If necessary, we must be prepared to increase our own skills—either through formal coursework or individualized training from vendors—to remain relevant in this new world. What we all must accept is that change is mandatory if we hope to function usefully in an electronic environment.

Conceptualization skills

Perhaps more important than specific technical skills is the ability to conceptualize how systems work. This is essential when identifying the appropriate level at which to inventory, appraise, and schedule e-records as part of the inventory process, as well as when devising strategies for their destruction. For example, a system that essentially functions as an index to linked images can be described in terms of series or document types because it replicates a series—or at least a grouping of document types—found in the paper environment.

But the concept of the records series, which is always a bit fluid, is chimerical in the electronic world, and the concept itself might require rethinking. While dealing with a large financial system, our thinking began with paper: we expected purchase orders, monthly purchasing reports, and final fiscal reports. Instead, we discovered one large, integrated recordkeeping system that encompasses ten or more series from the paper world. Essentially a large relational database,

the electronic system requires multiple departments to work together to create and manage the records. The system has user views that replicate paper forms, but these are simply data entry screens and access tools. Reports are not generated from them, and so the data is never separated into individual series. The electronic system has reconceptualized financial tracking, collapsing what had been multiple paper records series into one electronic series with a broader functional scope. This represents a change of little importance to the office workers using the system, but one that we as records managers and archivists recognize as a momentous transformation.

Appraisal and scheduling skills

We need to rethink appraisal and retention scheduling. We might find that records in an imaging system still conform to the old divisions by series. In large databases, however, data can be so interconnected that it may be necessary to retain everything for the maximum retention period, much as we would retain hardcopy records of varying retention periods that are boxed together. These simple scenarios demonstrate that born-digital records are more likely to challenge archival concepts than are digitized records. This project has also uncovered problems with our traditional general schedules. When a recordkeeping system never comes to an end, we no longer have the opportunity to use retentions such as "6 years after a volume is filled." Instead, we are forced to reconsider longstanding methods of defining retention periods and to revise our retention schedules to conform to new realities.

Ideally, archivists and records managers should work with IT professionals to design systems that accommodate existing records schedules, but this rarely happens in today's work environment. Until it does, we need to anticipate and develop a strategy for those governments that retain records beyond their prescribed retention periods. This retention solution has become increasingly common as storage capacity and access rates have increased, but it has many drawbacks. Keeping everything, plainly, nullifies the concept of archives, which holds as a truth that some few records are worth keeping forever. In particular, we should address the increased number of access requests—and associated risks—involved with retaining a constantly growing collection of records that may have to be produced in response to Freedom of Information requests, audits, and court orders.

In some instances, electronic recordkeeping makes retention easier. It can unite records that in a paper environment are maintained by two different departments, thus making it easier to manage conditional retentions. For example, it makes possible the timely destruction of purchase data at a specified period after the entry of payment data relating to that purchase. Learning to take advantage of these new possibilities in the retention process is another skill we need to develop.

Rethinking destruction

We need to determine how to certify that records have been destroyed when destruction is difficult to define and to verify absolutely in the electronic world. Deleting a set of records on a

server, for instance, neither obliterates those records nor addresses their duplicate stored on other computers or backup tapes. In the electronic world, the simple act of destruction has become more complex and less absolute at the same time. The profession still needs to develop methodologies for handling this otherwise mundane action.

Are indices integrated with a set of electronic records always part of that records series? Under what circumstances would an index to a series of records not be part of that series? Simply: When is metadata something different than part of the record itself?

How do users conceptualize records within complex records systems? Should archivists and records managers study these perceptions to identify new ways to segregate records into manageable "sets"? Is the concept of the series becoming obsolete? Is this consideration different for general records management activities than for archival activities?

How can archivists better manage records in electronic systems that focus on the now, reach back into the recent past, but ignore preservation?

How can archivists and records managers protect users and organizations from the pitfalls caused by the indefinite retention of some electronic records?

3. *Acquisitions: Assessment, Scheduling, and Transfer of Public Affairs Records*
Timothy Pyatt, Duke University

Public Affairs operations at institutions of higher education are rapidly abandoning paper as a medium for disseminating press releases, news stories, and campus promotional materials. Archives must develop plans to acquire and preserve these electronic records as paper surrogates are no longer produced. The Duke University Archives has found that while its "traditional" records scheduling and accessioning methods have worked to manage the intake of these records, "traditional" processing and access methodologies have not proved effective. New skills are needed to facilitate the transfer of electronic files and to assess file format longevity and authenticity. Reevaluation of processing procedures and the viability of traditional finding aids for electronic records needs to occur.

Scenario

As the official repository for the records of Duke University, the University Archives identifies, acquires, manages, and preserves University records of enduring value, regardless of format, and makes them available for use. Founded in 1972, the Archives holds over 11,000 linear feet of administrative records, campus publications, records of student groups, and selected alumni collections and faculty papers. The Archives has a staff of five full-time employees—University Archivist, Reference Services Archivist, Archivist for Student Life and Technical Services, Records Services Archivist (records manager), and an Archival Assistant. Graduate interns and student assistants assist with processing of collections. While the Archives maintains its own website, encodes and loads EAD finding aids, and does some limited digitization, IT support and server maintenance for the Archives comes from the Library's IT office.

The University Archives staff have spent substantial time with Duke's Division of Public Affairs staff discussing how they manage electronic records and have worked to capture that information in records schedules. Historically the Archives have received biographical files about faculty and staff, news releases, the campus staff newspaper, and other campus promotional materials from Public Affairs in paper and analog media formats. As with many campus offices, Public Affairs has started to phase out paper publications and no longer distributes press releases and news stories in the paper form. Our records manager created a retention schedule for public affairs, news, and communications records, which included the following text describing electronic records:

> Public affairs, news and communications offices rely almost exclusively on electronic or born-digital records created with word processing, spreadsheet, electronic mail, website authoring, or database programs to carry out their business and activities. While the all-encompassing term "records" includes any recorded information stored on any medium, the guidelines attempt to recognize that a series may contain paper and electronic records. As electronic records are created, managed, and stored

> throughout their life cycle of usefulness, these guidelines should serve as a general strategy for identifying basic retention needs for different sets of information. http://www.lib.duke.edu/archives/rm/PubAffairs-Final.htm

In the summer of 2004 the Archives worked with Public Affairs to transfer the bulk of their non-current paper records with archival value to the University Archives and started discussion regarding the transfer of non-current electronic files with archival value. Public Affairs maintains the website for Duke's president, and when President Keohane retired in 2004, the staff wanted to take her site "off-line" but still preserve it. Keohane was the first Duke president to have a website. The Archives acquired the files and placed them on a server dedicated to storing archival electronic records. The next significant transfer occurred in 2005 when Public Affairs migrated its website to a new content management system (CMS), rather than migrate all press releases and news stories in electronic records creators.

All of the Public Affairs electronic records transfers received were non-current and unrestricted (open for public use). As soon as an access/delivery system can be developed, these could be made available for public use. As a test, we created PDFs of speeches by former President Keohane from her website and made them available online (http://www.lib.duke.edu/archives/history/ keohane-spchndx.htm). While this works for the 100+ Keohane speeches, we do not see it as a good solution for the 18,000+ press releases and news stories.

In this case the Archives acquired records through the use of FTP, discs, and web harvest. For future transfer of news releases, we have discussed using a yet-to-be-determined automated process for capturing selected content. The records transferred in 2004 and 2005 were accessioned after receipt and stored on a server in directories by accession number. This parallels the process used for paper records and has been effective to date. While the metadata for the files were inconsistent (i.e., no standard naming conventions), the files were organized in directories (folders) by year and by series (i.e., all of the 1999 press releases in a single directory). Authenticity and version control was more difficult to determine as a 1999 press release may have been reused in 2003 and resaved with a 2003 date, but still be in the 1999 folder, making it impossible to determine what exactly was released in 1999. Within a single year there is not a consistent file naming convention; each file must be opened to determine release and content. This can be challenging as a variety of word processing software in multiple versions have been used for the files.

While the content and handling of the press releases, news stories, and speeches transferred largely mirrors the content and process for their analog predecessors, appraisal and transfer of the digital media files created by Public Affairs have proved far more difficult. These files include interviews with faculty experts, Duke-produced news stories, promotional pieces, and clips associated with a news story or press release. Third-party media files, such as public TV interviews with campus leaders and experts, are also among these files. The metadata for these records are problematic as standard naming conventions are not used and often no direct linkage to the story or press release exists. However, the press release has an embedded link to the

streaming version of the media file, giving a connection from that side of the content. Two or more versions of the media clip are often available—an uncompressed version (with a file size of over 400 MB) as well as MPEG4, Real Media, and/or QuickTime versions (file size in the 40 MB range). The process for review and transfer of this content is still in the planning stages.

As long as secure server space has been available, the intake of records has worked well. Using the same process traditionally used to accession paper-based records has allowed us to gain initial control over the public affairs content received. Source files were transferred to the Archives and then placed in directories by accession number on the dedicated digital file server shared with the Library's special collections. Access is limited to Library IT staff and selected staff from Archives and Special Collections. The server follows the Library's standard back-up routine with back-ups run daily. This process is roughly analogous to our process for paper records. We accept the records, assign a unique accession number, and locate a secure storage area. The main difference is the storage area for the unprocessed records—directories on a server instead of shelves in the stacks.

The process of creating the accession record itself also parallels that for paper records. The Archives creates accession records using the cataloging module of ALEPH, the Library's integrated library system. Accession records are created as "suppressed" catalog records using a mix of MARC and locally created fields. The administrative data of the transfer, such as office of origin, type of records, volume, and location, can still be captured. The volume is listed as a number of electronic files instead of boxes and the location is the server name rather than a stack location, but otherwise the content of the record is the same as if the files were paper-based.

The issues of file management before transfer are not unlike records management challenges posed by paper files. Consistent use of filing names and schemes have been a source of continuous training in the realm of paper records, so it is not surprising that this continues to be an issue with digital records. The concerns about file authenticity are more vexing; in the paper era if a 1999 press release was reworked to be reissued in 2003, the original paper copy of the 1999 press release would not have been directly rewritten on the original paper and then re-filed with the 1999 releases. The 1999 original would have remained in the file and the "new" 2003 version would have been created and filed with 2003 releases. The ease of rewriting an existing "active" file in the digital era, even when it is the "copy of record," has significant authenticity issues for archivists that must be better communicated to the records creators.

The processing of digital records and our ability to provide access has also been a challenge, and one where the analog model has not matched as well. For paper-based records, processing and providing access has entailed describing the records to appropriate level, creating and encoding a finding aid, loading the finding aid and catalog record on the website, and then making the collection available in the reading room for the Archives. While the processing and finding aid can still be done, reading-room-only access to digital content is not the desired or expected access. For non-restricted content, users expect a web-based interface accessible from

any location with Internet access. Archival finding aids do not generally offer the same level of searchable metadata that users expect from online digital records. While the finding aid may offer searchable folder lists for paper press releases, users want to keyword search the full text of the actual items. For the Archives to provide such access, each of the 18,000+ press releases and news stories would have to be migrated to a standard format (such as searchable PDF), loaded on our website, and made available through a search engine. This item-level work has been done for selected content (the previously mentioned presidential speeches), but was too labor intensive to use for the larger volume of files.

New Skills

What technical skill issues exist for archivists when assessing, scheduling, and transferring electronic records created by public affairs?

- Ability to identify content and formats: just as archivists can presently assess paper and photograph formats and determine their longevity, we need to develop the same skills to determine the longevity and viability of digital file formats. We must understand the preservation and access issues associated with varying file formats.
- Knowledge of file transfer protocols: what are the best and most efficient methods to acquire digital content without losing context or authenticity? We need to communicate and collaborate with appropriate campus IT staff so they will understand our desired goals and outcomes.
- Knowledge of software that can work across formats to appraise content: this is a major issue as digital content can be in file formats ranging from old versions of Word Perfect to HTML to XML. Software such as Quick View Plus can allow one to view a file without altering it.
- Application of appropriate metadata in order to retrieve content: this can range from basic metadata in the accessioning process to potentially full text access for "processed" materials.
- Ability to document and preserve file authenticity as part of the transfer: do the files need migration? Do multiple versions exist?
- Ability to plan for and manage appropriate server storage space using language that campus IT professionals understand: just as in the past we needed skills to plan for, justify, and request additional stack space for collections storage, we need to develop skills to make similar compelling requests for server storage space.

4. *Guarding the Guards: Archiving the Electronic Records of Hypertext Author Michael Joyce* Catherine Stollar, Harry Ransom Center, University of Texas at Austin
Thomas Kiehne, Fosforus

In 2005, the Harry Ransom Center at the University at Austin acquired the fonds of hypertext author Michael Joyce. The major emphasis of the Ransom Center's collections is the study of literature and culture in the late 20th and early 21st century of the United States, Great Britain, and France. Michael Joyce's groundbreaking work in hypertext poetry and fiction make his papers a desirable addition to the Ransom Center holdings.

The Michael Joyce Papers are mostly composed of electronic records with an additional 60 manuscript boxes of paper-based materials This is the first mostly electronic archive the Ransom Center has acquired and new strategies for preserving digital content were employed. This case study discusses the techniques and skills utilized to preserve the electronic records of Michael Joyce as a model for processing future digital manuscripts at the Ransom Center.

Scenario

Established in 1957 by University of Texas Vice President and Provost Harry Huntt Ransom, the Harry Ransom Humanities Research Center at The University of Texas at Austin incorporated a strategy for collecting older rare books and manuscript collections with a new initiative to collect literary, photographic, and theatrical works by modern artists. Some of the authors whose works are included in the Ransom Center's collections are Norman Bel Geddes, Don DeLillo, T.S. Eliot, James Joyce, Ernest Hemingway, Norman Mailer, D.H. Lawrence, Ezra Pound, Anne Sexton, Isaac Bashevis Singer, and Tennessee Williams. Michael Joyce's work as perhaps the most influential hypertext poet and author fits nicely into the Ransom Center's contemporary author collecting policy.

Our case study to preserve Michael Joyce's digital manuscripts resulted from collaboration between the School of Information at the University of Texas at Austin and the Harry Ransom Center. Three students, Thomas Kiehne, Vivian Spoliansky, and Catherine Stollar, from Dr. Patricia Galloway's Problems in Permanent Retention of Electronic Records course offered at the School of Information, undertook a semester long project to develop a strategy for archiving an initial accession of electronic materials saved on 371, 3.5 inch floppy disks (totaling 211 KB) from author Michael Joyce. Upon completion of the project, a second accession of electronic and paper-based materials, including the contents of three hard drives (totaling 8.38 GB) and 60 manuscript boxes, was acquired by the Harry Ransom Center and is currently being processed by staff archivist Catherine Stollar according to the strategy developed during the class project. Our case study discusses strategies for file recovery, migration, preservation, arrangement, and description developed working with both accessions of Joyce's materials. The electronic records are currently maintained in a DSpace repository administered by the School of Information; however, in the future, the Joyce records will move to a DSpace repository controlled by the Ransom Center and the General Libraries of the University of Texas.

The lack of digital archivy skills among Ransom Center staff provided the impetus for the Ransom Center to partner with the School of Information on the Michael Joyce Papers. Although the Ransom Center employs talented archivists and IT professionals, no staff member possessed the skills necessary for archiving digital manuscripts. The Ransom Center sought advice from Professor Galloway and agreed to use the Joyce materials as a case study in the Problems in Permanent Retention of Electronic Records course.

Some audio and video migration preservation projects were already in progress at the Ransom center in the Department of Photography and Visual Collections, but there were no concerted efforts to preserve born-digital manuscripts. Policies and procedures for migration of audio and video content to new media were unsuited for born-digital manuscript preservation, and policies for preserving digital manuscripts were inadequate to capture the complete behavior of the original digital record. Previously, the few electronic manuscripts and correspondence already in the Ransom Center's manuscript collections were printed and organized in boxes like paper records. Because digital records are entirely unlike paper-based records, a preservation strategy based in printing records preserved very little of the original document. Electronic media were saved, but researchers were prevented from viewing original disks and no access copies were created.

The main component of our preservation strategy is to ingest electronic records and associated metadata into an institutional repository. DSpace, created from a joint project between MIT and Hewlett-Packard, is the institutional repository we used and will continue to use for electronic record preservation. At the heart of DSpace, like most open archival information systems (OAIS), is a database populated by individual digital objects supported by content, context, and structure metadata. We used DSpace, instead of FEDORA or another institutional repository, because it was already established as the repository of choice for the School of Information. Although we had issues with the web user-interface for ingesting, viewing, and accessing materials within the repository, we plan to work with a talented iSchool student with Java programming skills to make our installation of DSpace more user-friendly.

Partnerships were key components to our case study's productivity and success. The initial group of students participating in the first part of the case study each represented a different skills background. Thomas Kiehne brought a wealth of information technology skills to the project, including programming and operating systems knowledge. Catherine Stollar shared her knowledge of archival theory and practice during the case study. Vivian Spoliansky viewed the case study through the lens of preservation and shed light on aspects of authenticity and desired levels of service for object preservation. Working with a variety of subject specialists on the project enabled participants to learn key skills that will be useful on future digital record preservation projects.

Processing as Digital Archeology

One of the more unique aspects of this project involved the processing of 371, 3.5 inch floppy disks that contained the digital objects of the first accession. The floppy disks were mostly from

the Macintosh "classic" era, dating as far back as the mid to late 1980s. The assumption at this stage is that the original storage media is not stable or reliable and the information that the floppy disks hold must be moved quickly and efficiently.

At the outset of the project, we had only a general idea of the process of moving the digital files from the source media to a repository, and as such, we could not express specific requirements for software tools and utilities that might be needed. Little was known about what to expect in terms of specific technological issues or challenges. In order to minimize project overhead in terms of time and resources, we desired to use only readily available open source, shareware, or freeware tools to assist with the extraction process. This approach allowed us to assess the suitability of tools that are currently available and their ability to interoperate. In the absence of suitable free tools, we intended to find commercial software or create our own programs or scripts to perform the required tasks as we identified them. In the course of processing the first accession of disks, we quickly elucidated a more detailed procedural framework that can be abstracted and applied to future projects.

The general process implemented during the processing of the disks is as follows:

1. Receive and identify physical media
2. Catalog the physical media
3. Copy files to newer physical media
4. Perform initial file processing
5. Create an item-level index of all recovered files
6. Create and process working copies of all files while retaining the original bitstream copies

Technical metadata is collected at each step in the process not only to facilitate the work in progress, but to support provenance and authenticity. Each operation performed on the bitstream—every copy and access—provides the opportunity for inadvertent loss or alteration, so careful recordkeeping is as essential as careful handling. Additionally, all personnel involved in processing must thoroughly understand the procedures involved in order to prevent duplication of effort or discontinuities in results. In many cases, software can automate these processes, thus reducing the chance of errors, but the extent to which software can mitigate such risks is limited by the assumptions made by the creators of the software and how well the personnel making use of this software understand these limitations.

Given that time was of the essence, we opted to use text entries in Microsoft Excel spreadsheets to create the initial disk catalog and the associated metadata. This approach allowed us to leverage existing proficiency with spreadsheets and the availability of the software to eliminate the time needed to create a custom database application or to learn project management software. Unfortunately, the absence of relational or workflow aspects in the spreadsheet format made us vulnerable to recordkeeping errors, making quality control a primary concern.

The copy functionality of the computer operating systems involved were sufficient to perform the movement of digital files from floppies to hard drives and removable media. Unfortu-

nately, the differences between Macintosh and Windows in the management of file system metadata became significant. Creation dates are handled differently in these two operating systems so that a copy made in Windows takes on the date of the copy operation, not the creation date of the original from which it was made. Additionally, file system metadata for Macintosh files are stored as separate, invisible resource forks that are notorious for becoming corrupted. As a result, we often could not trust the dates ascribed by the operating system and had to refer to external resources, such as Michael Joyce's curriculum vitae, to confirm or provide date metadata at a later time. Issues with Macintosh resource forks also affected file downloads from DSpace after ingest.

At many points during the processing, we encountered technical difficulties in the form of file or disk errors. These errors can occur for a number of reasons, including damaged media, exposure to magnetic or other hazards, dirty data surface areas, and so on. Several attempts were needed to overcome a copy error caused by a dirty surface area, resulting in a suggestion that a drive cleaning kit be used periodically to prevent build up of debris on the drive head. For other errors, it was necessary to have available software utilities that can attempt recovery of files from copying errors. Windows provides such capabilities within the operating system (e.g.: Scandisk), but Macintosh does not. For our purposes we were able to discover an older version of a commercial program, Norton Utilities, which allowed us to recover many files that could not be copied initially. Virus checking was also a preeminent concern. Errors and crashes must be met with persistence as they are often surmountable, which implies at least a minimum degree of technical knowledge.

In moving the digital files to other media, we created a file system hierarchy that mimicked the physical arrangement of the disks. Such hierarchical arrangement allowed us to use file system tools to generate some of the metadata automatically. There are a number of freeware, shareware, and commercial applications for Macintosh that will catalog a file volume and produce reports. We used a shareware utility called CatFinder[59] to index the copied files and export a report to a delimited format that was imported into Excel. This report formed the basis of our item-level metadata, including fields for file name, file size, kind (document or folder), Macintosh file type (analogous to the Windows file extension), Macintosh creator code, creation date, and modification date. To this basic report we added a comments field for use during appraisal and to collect technical notes.

MD5 file hashes were also generated for each file. Having an MD5 hash for each file allowed us to do two important things: to identify and/or eliminate redundant files, and to support provenance auditing during the repository ingest process. A freeware PERL application called Integrity[60] automatically created MD5 hash calculations and exported the results to a delimited text file. Unfortunately, integrating the MD5 hashes into the CatFinder index was not trivial

[59] http://www.mindspring.com/~shdtree/newsite/id9.html

[60] http://therockquarry.com/integrity.htm

due to differences between the two applications in file name recursion and handling of hidden files.

Having created a unified index of file system metadata, augmented with processing notes and MD5 hashes, we were able to more accurately assess the extent of the digital files and facilitate arrangement and appraisal. Unfortunately, the index was in no way tied to the digital files and presented us with a significant information management problem. For example, any movement of files was not automatically noted in the index, nor was any change or deletion in the index reflected in the file system. We can envision a workflow-oriented system that stands between the file system and a metadata database that would greatly increase the speed and reliability of processing large bodies of digital documents.

Arrangement[61]

After recovering most of the unique digital content from the first accession of floppy disks, we began the process of archival arrangement. In the beginning, we asked ourselves some questions. Can and should digital files be arranged like paper-based records? Should we heed traditional archival arrangement practices or follow newer theories of arrangement based on item-level metadata? Do electronic records have a natural hierarchy that can be expressed in a traditional arrangement? Should physical housing for digital materials be kept? If so, where? Our answers to these questions are not definitive, but we came to a compromise incorporating basic tenets of archival theory with features of on-demand, flexible file arrangement using item-level metadata.

A number of digital materials within the archives, including e-mails and published articles, had a paper-based counterpart, demonstrating that Michael Joyce created both digital and analog records while performing the same activities. Both formats of records were created synchronously, and at an institution like the Ransom Center that preserves not only works that have influenced the arts and humanities fields, but also preserves the context in which those works were created, we determined it would be desirable to reflect synchronous creation in the arrangement. We did not originally understand relationships between Joyce's digital and paper materials because the first portion of the case study dealt only with electronic records from the first accession of floppy disks. We initially arranged the files into five series: Works, Academic Career, Correspondence, Storyspace, Third-party Works, and Personal. After surveying the paper-based materials and the second accession of electronic materials, we had to alter our original arrangement to include the newly accessioned materials. The final arrangement created is Works and Related Materials, Academic Career, Correspondence, Storyspace, Journals and Appointment Books, Personal, Works by Other Authors, and Published Materials.

Institutional repositories like DSpace can facilitate digital object arrangement into our specified series by using the community, sub-community, collection, sub-collection, and item-level hier-

[61] More detail about this part of our project can be found in a forthcoming article about processing the Michael Joyce Papers in *Provenance*.

archies. DSpace's hierarchies relate to traditional archival hierarchical levels: communities equate to archival *fonds*, sub-communities to series and sub-series, collections as other layers of granularity within a series, and item-level entries relate to digital objects. In an additional level of granularity, items composed of multiple sub-components or related files, e.g., websites with multiple linked HTML files, can be ingested as bundled files.

After determining how to arrange the paper and digital materials, we decided on a way to arrange the physical housing (jewel cases, magnetic media, paper holders, plastic cases, etc.) from Joyce's electronic works. Previous policies and procedures at the Ransom Center dictated that electronic media should be physically housed in Hollinger boxes separate from the rest of the paper-based materials. This separation policy apparently arose out of concern for potential damage to other materials caused by degrading electronic media and as a way to limit access to the electronic materials by researchers. No studies on electronic media degradation have found any instances of off-gassing or other damaging effects of filing electronic media with paper-based materials, so we determined that physically integrating paper-based material and digital media would be the best policy for physically arranging the Michael Joyce Papers. The Ransom Center will still limit access to files saved on original media because researchers will have access to the files via DSpace.

Although we integrated Joyce's digital objects into a functional group arrangement similar to his paper-based records, we also took advantage of the flexible, non-linear nature of digital object arrangement by enabling on-demand, user-controlled arrangement by item-level metadata. Preservation of digital objects depends on item-level metadata used to document, migrate, emulate, and preserve the objects. Item-level metadata recorded for preservation in DSpace's database also enables flexible arrangement of digital objects. Digital arrangement allows archivists and users multiple options for organizing objects depending on the parameters set by the user interface, such as file name, title, author, date created, subject, or other metadata element. Arrangement is limited only by the skills of the programmer developing the user interface and the precision of metadata recorded for each object.

Arrangement was also affected by how we ingested objects into DSpace because our method of ingest affected what metadata fields were included. Although manual metadata assignment of all files within the Joyce archives was laborious, certain metadata fields were impossible to record automatically. Content metadata, such as *subject* and *title of work*, had to be entered by hand because automatic tools to accurately extract content were not available.[62] We found it difficult to use file names within the archives to associate files with published titles because the file names were not specific or standardized.

[62] Literary text comparison tools designed for use with small numbers of digital works were not sufficient for our large collection of files. Apparently text-mining tools could serve our purposes by comparing large bodies of records with each other. We have not utilized any text-mining tools to date.

We incorporated methods for traditional archival arrangement and strategies for on-demand item-level arrangement while processing digital objects within the Michael Joyce Papers. Together, both methods allow users to browse records according to functional series and create new arrangements based on any metadata available for individual objects.

Challenges

In addition to the challenges we encountered in developing a strategy to preserve Joyce's text and graphic files, we faced unique challenges associated with preserving Joyce's most influential creative works written using specialized software called Storyspace. The Storyspace format type, created by Michael Joyce, Jay David Bolter, and John B. Smith, presented (and continues to present) challenges for media migration, ingest, and file use. Hypertext works written in Storyspace are composed of multi-faceted texts linked by guard fields (words within texts that enable direct links to other nodes, usually under specific conditions) and can only be viewed using Storyspace software. To complicate matters, we originally thought the latest version of Storyspace was backwards compatible and could read works written in the first version of Storyspace. Unfortunately, this is not entirely the case as older Storyspace documents do not degrade gracefully. For example, the text from files written in Storyspace 1.5 can be read in Storyspace 2, but the individual nodes and links are missing, making the Storyspace 2 rendering of an older work vastly different from the original.

New Skills

- A thorough grounding in the various operating systems. The profusion of technical difficulties and operating system inconsistencies required an intuition about the various platforms that can only be gained by direct experience. While processing digital files it is essential to have an understanding of the environment in which they were created. Computer literacy in more than one platform and with networked environments will be ideal skills for archivists of the future.
- A basic understanding of the structure of digital documents. Knowing how a digital file is created and stored, including such basics as the difference between binary file formats and textual formats such as ACSII and UTF, helps provide an understanding of what happens during processing. Furthermore, an intimate knowledge of the types of formats and how they might be identified (e.g., file type extensions or creator codes), accessed, and converted is essential. Understanding digital formats offers clues to where to find item-level metadata (e.g., document properties embedded in word processing files, ID3 tags embedded in MP3 audio files, etc.) and suggests migration paths for long-term preservation.
- Proficiency with and trust in new tools. Familiar means of handling physical documents are not present with digital documents. Software tools and operating systems augment the functions of our senses in the digital world and mitigate some of this loss, but not completely. Integrated toolkits and processing systems are needed and must be developed so that they can be trusted to conform to the expectations of archival practice.
- Establishment of new workflows and procedures. The intangible nature of digital information makes documentary evidence crucial to processing. Many institutions have

established procedures for document processing, including audio/visual materials, but these cannot be assumed to be sufficient for digital objects. Operating systems alone cannot document processes, so new systems that function according to sound processing policies are necessary.

- Ability to monitor current trends in digital preservation including metadata standards, crosswalks between encoding standards, available tools, storage systems, file format repositories, national and international research initiatives, user expectations, and published best practices guides.
- A thorough understanding of traditional archival theory and practice. Archivists who work with digital records should be able to extrapolate traditional theory and apply it to electronic record preservation, but must be flexible enough to create new standards for archival practice. What we do as archivists will change (practice), but why we do it will not (theory).

5. *The Eyes of Texas: What can archivists learn from working with a digital institutional repository?*
Patricia Galloway, School of Information, University of Texas at Austin

Establishing even a small-scale archival digital institutional repository calls on every skill archivists can muster, both inside and out of the (technological tool-) box. In this case study we discuss one category of collections, faculty papers, as acquired and added to the departmental institutional repository created in 2005 for the School of Information, University of Texas at Austin. The nature of such collections requires of the digital archivist technical skills ranging from retrospective digital archaeology (presaging migrations and emulations to come) to prospective records creation management (asking archivists to manage an ongoing interaction among creator, records, and repository).

Origin of the project

We wanted to establish a departmental institutional repository to serve University of Texas School of Information faculty, staff, and students by providing a secure and persistent environment to preserve and expose faculty scholarship, to archive scheduled departmental administrative records including faculty committee work, to preserve departmental digital productions (websites, tutorials) for historical reasons and as teaching resources, and to provide a repository for the preservation of the preprofessional work of students. We felt that the repository, if managed to instantiate archival standards on various levels, would exercise many of the skills needed for student archivists to learn to work with digital records, while additionally becoming an asset to other interests in the school.

How does this project address the issue of "new skills"?

Skills in the analysis and facilitation of the social ecology of record creation and use are vital to the management of all records, but they are especially important in the case of digital records because new environments have led to new uses and communication practices. Like all electronic records projects, however, this one also calls on an amazingly broad panoply of technological skills, because it demands not only that we understand the objects themselves and their formats, but also that we develop a sophisticated understanding of the system environment in which the objects must be processed and the repository environment in which they must be placed and managed. At the present stage, the focus is on pre-processing of records to be ingested into repositories and metadata capture to support future preservation activities. I anticipate that our students will be struggling in the future to solve the problems raised by the materials we ingest today, but if we wait to ingest until we know everything, there will be nothing for them to struggle with.

Background of the case

Since 2003 I have been working with graduate students at the University of Texas School of Information (iSchool) on problems of archival permanence for digital objects using the open-source DSpace repository software as a testbed for our thinking and practice. We have worked

on a design and policies for a centralized e-mail repository for Texas state government (2002),[63] a website repository for the School (2003, 2005 to 2006),[64] preservation of papers and publications of iSchool faculty (2005), administratively created materials belonging to the iSchool (2005 and 2006), and special projects not directly involving the iSchool except as temporary digital repository, such as policies and practices for the permanent preservation of online scholarly journals and anthropological field notes for the American Anthropological Association (2004 and 2006)[65] and the Michael Joyce Collection hosted for the Harry Ransom Humanities Research Center (see Stollar and Kiehne project).

In 2005 we undertook the formal establishment of a digital institutional repository for the iSchool, beginning with faculty records and the recovery and preservation of the School's website in its historical incarnations. The faculty records subset of this task we called the "Eyes of Texas," in reference to the UNC-SLIS "Minds of Carolina" project for collecting the digital records of retiring faculty at the University of North Carolina at Chapel Hill. We used version 1.2.1 of the DSpace repository software. Theoretical principles and grounds for practice for all the procedures not instantiated in DSpace were adopted from the Open Archives Information System reference model and broad readings in the work of digital projects worldwide, with special reference to big projects like InterPARES and NARA research and special thanks to the industrious Dutch and Australians.

For the purposes of this case, I will concentrate on the "Eyes of Texas" project from 2005, although developments this year, with projects just completed, have greatly expanded some findings and will be mentioned as appropriate. I received a small grant sponsored by Ingenta and granted by the ALA Research Round Table to focus on the role of these collections in the institutionalization of a departmental repository, and as a result we have followed up on these collections and the institutionalization process this year. The creator-participants in the "Eyes of Texas" project were four faculty members, three at the end of their faculty careers and one (our dean) who is at active mid-career. We asked them to tell us what they wanted to preserve as emblematic of their work and potentially make available in this way to both our school and to broader university and public audiences, and their perspective was vital to our understanding of the contexts of record creation and use. We built on work done in the same seminar in 2003 to develop repository policies on levels of service and support for different formats and types of materials and intellectual property, privacy, and confidentiality concerns, but we added to that work as additional issues emerged in the process of actually committing to support real materials.

---

[63] Marlan Green, Sue Soy, Stan Gunn, and Patricia Galloway, "Coming to TERM: Designing the Texas E-mail Repository Model," *D-Lib* 8, no. 9 (September 2002).

[64] Anne Marie Donovan, Maria Esteva, Addy Sonder, and Sue Trombley, *Proposal for Establishment of a DSpace™ Digital Repository at The School of Information, University of Texas at Austin,* May 10, 2003, available at http://hdl.handle.net/1721.1/1273.

[65] See documentation filed in Portal Archiving Study Project, https://pacer.ischool.utexas.edu/handle/1721.1/45 and in the ASSC Special Project, https://pacer.ischool.utexas.edu/handle/123456789/1333.

What did we do with electronic records?

- Understand the creators' relationships with their records. Our first concern was to elaborate on typical inventory practice and emergent interests in self-representational aspects of private records to understand how the collection of digital materials each faculty member offered also mirrored the history of the computing environments each has experienced.
- Establish submission agreements with creators incorporating intellectual property and preservation commitments. As we all by now well appreciate, a typical SIP agreement goes far beyond the usual donor agreement (even one that anticipates accretions), especially in that it must remain a dynamic document.
- Capture the features and structure of the environment in which the records were created and used. In most cases we were able at least to control this information partially through historical institutional knowledge of the IT environment of the iSchool.
- Capture and catalog the technical characteristics of the records themselves, capturing and providing object-level metadata where feasible.
- Design a repository structure for the records to inhabit, incorporating collection-level metadata. Here our archives students had the opportunity to bring their skills to bear on digital collections.
- Preprocess the records prior to ingest. This preprocessing is not built into DSpace, so we followed best practices to set up workflows that would preserve intrinsic metadata intact while carrying out tasks like copying and testing for viruses.
- Ingest/accession the records, adopting and adapting the DSpace workflow.
- Prepare access versions of some formats.

What did we learn about working with electronic records? What skills did we need?

- Complex problems of format/operating system interactions (like the Apple/Macintosh resource fork) meant that we could not even capture records and move them into a preservation environment without a sophisticated knowledge of both the creation and preservation digital environments. Further, because even in the case of the active faculty member, files to be dealt with had been created with several different versions of software with some time depth, we had to proceed carefully so as to preserve the object intact for archiving before we moved to effect access. These complex issues meant that it was vital for us to consider from the outset which significant properties of each collection we needed to be able to preserve (beyond foundational bitstream preservation) and how we planned to render them on access.
- In one case, a faculty member wanted to include postprints but did not have any digital versions of older publications, so the student team set up a mini digitization project to satisfy this requirement. Here we instantiated digitization best practices as carried out by the UT General Libraries' Digital Library Services department.
- In supplying keywords to assist in retrieval, we found that LCSH were both too formal and not detailed enough to be of much help with highly technical articles, especially when compared with new access methods like tagging and full-text search. Accordingly, one student undertook an independent study to experiment with using data mining techniques on the texts of submitted publications for extracting keywords from the deposited texts themselves in order to

establish a controlled vocabulary for subject description in that collection. This year other projects have made use of this procedure.

- Diagnosing obscure file formats and capturing environmental metadata required the use of tools as various as hex editors, directory listers, metadata extractors, legacy viewers, and web crawlers. Where these tools were open source and open access, we collected and archived them to form the nucleus of a working digital tool shed.
- Conventional archival description is aimed at the aggregate levels of archival collections rather than the granular individual object level. This means that for paper archives there is in fact—apart from ill-codified practice in the making of calendars and the like—very little "best practice" for metadata allocation at the individual object level. DSpace itself is designed to capture a certain amount of technical metadata on ingest, but the depositor is required to provide a range of Dublin Core metadata chiefly aimed at resource discovery. We are still in the process of wrestling with appropriate descriptive practice at this level and plan to delve into the consistent structuring of collection-specific metadata this fall in a class on metadata, now that we have worked through several solutions.
- Decisions favoring access were taken to produce more accessible "use copies" through conversion from the original objects (for example, JPEGs derived from TIFFs). It is clear that for larger-scale projects than we have yet tackled, a migration-on-demand tactic will be more appropriate, which will mean the provision of a digital workspace and appropriate tools for users to wield.
- Intellectual property issues proved to be quite complex. We made use of the online SHERPA registry (although in early 2005 it was still very much in formation and did not cover all the periodicals/publishers we needed) to determine self-archiving policies of publishers for published materials that we would instantiate in providing access copies.
- We referred to archival ethical practices regarding privacy to deal with a large quantity of e-mails, which we decided to preserve as an encapsulated object to remain closed for the present and to be used for research on e-mail use in the medium term only with permission.

How did working with these materials differ from working with analog materials?

- Obviously the materials themselves are mediated: it *mattered* whether we looked through a Windows, Mac, or Linux lens. Students learned to *see* their mediating tools and not just *use* them. This made it clear that there is a need to establish and document a consistent preprocessing environment for handling digital objects.
- We were much more dependent on creators for learning about the records themselves and the environment in which they functioned than archivists of analog records usually feel they need to be. In addition, simply because most electronic records are being obtained much closer in time to creation than are paper records, our creators were still very much alive and concerned with their materials. This meant that our practice, like that of collecting archives, had to be much more participatory and had to take much more account of the creators' wishes.
- We had to be concerned from the beginning with intellectual property issues; the fact that digital records have the potential for nearly instantaneous availability, together with the enhanced awareness of these issues in a university environment, meant that IP and access management had to be fundamental from the start to a degree that is quite foreign to paper en-

vironments where processing alone can take years and soften the problems attached to IP issues.

Postscript: Repository as Social Object

I was and remain interested in the problem of institutionalization as crucial to permanence in digital archiving, and I believe that institutionalization depends vitally on local commitments and participatory development. Campus-wide repositories are receiving all the promotion that university public relations departments can offer, but in many cases they have found it difficult to attract materials. As a result of our work on all the projects mentioned above, we have developed the idea of a campus federation of departmental and special library digital repositories with a central repository located in the general library institution. This layered concept, which we have discussed with digital library principals in the UT General Libraries, answers both to the need for secure dark-archive storage of original bitstreams to guarantee authenticity and to the now-accepted standard requirement that any trusted digital repository must provide a succession plan in the case of its dissolution. This concept, however, needs to be proofed in practice, and that is what we are engaged in now.

We feel that it is particularly important to capitalize on the strong sense of local ownership that develops within a community of practice such as an academic department, and to recognize that such an institution consists of several important stakeholder groups—faculty, administrative staff, IT staff, and students—each of which needs to be involved in the project as it progresses. The iSchool repository is being anchored in iSchool practice in several ways:

- We are beginning to archive materials that must be kept to a specific standard because they are administrative documents under an approved records schedule. This year we began with website postings that constitute the record copy of policy statements, and we anticipate working next year with the office manager to archive digital meeting minutes going back several years.
- Faculty collections are being monitored via ISI and Google Scholar for their effect on faculty scholarship impact, and preliminary results have attracted the interest of a second potential cohort of faculty member-donors, including Professor Loriene Roy, president-elect of ALA. Sharing scholarship has long been a goal of early adopters of repository technology, and even faculty who may want to post writings on their own websites are interested in archiving them in the repository.
- Among the faculty collections are learning objects and other materials which, because their creators are our own faculty, are relevant to the ongoing instructional activities of the school. These are the kinds of materials that are rarely collected by university archives, yet the digital environment permits them not only to provide expanded provenance for more conventional faculty output, but also to be deployed for their functionality both as relevant now and as examples of information objects for future studies.
- Close collaboration with the IT staff and webmaster have moved us closer to embedding the repository and its support in the ongoing IT service of the school. Because of the special security that the repository offers, IT staff are beginning to see benefits in what I would call

"archiving in the strong sense," to protect their investment in historical versions of the website, for instance.

- Student interest in depositing work in the repository to support future professional careers has manifested itself this year in the form of projects laying the groundwork for the ongoing capture of an online student publication in preservation and conservation and for the self-archiving of student portfolios.

Publicity, as always, has helped in addition to collaborative efforts and as such has become an organic part of the process. In September of 2005 students exhibited five posters about repository-based projects at the annual meeting of the Society of American Archivists in New Orleans, and one of them was honored with one of two prizes. Three of the projects produced papers that were submitted to *Provenance*, a journal about archival practice, and the work of the Michael Joyce project, which prepared and deposited the first digital collection accepted by the Harry Ransom Humanities Research Center, is now being featured in an exhibit, "Technologies of Writing," at the HRC. In addition, a brief report on the SAA posters was featured in the University of Texas-Austin alumni magazine, *The Alcalde*, and I was invited to give a presentation at the 2006 ALISE meeting on the use of an institutional repository for LIS teaching. Finally, I was granted tenure toward the end of 2005, and that fact, since it makes it likely that I will be involved with the repository for some years to come, has led to additional support for the repository, totally apart from the recognition of its use to faculty.

Other specific events, not uncommon or unavailable to others, had an impact as well in providing opportunities for the repository to demonstrate its usefulness. Because a redesign of the school's website was undertaken in fall 2005, the value of the repository as a secure archives for past designs came to the forefront. In addition, a new sequence of digitization courses, designed to serve digital library requirements, has made the necessity for a secure archival repository for such materials obvious, especially since I am involved in team-teaching the introductory course that stresses archival care for these derivative digital objects. During 2005 also, an online publication of the school's Kilgarlin Center for the Preservation of the Cultural Record, *The Cochineal*, was targeted as the first candidate for deposit of student work, and this project has gone forward this spring. Finally, student interest in learning about archival institutional digital repositories has grown and students are beginning to undertake individual research projects using the repository. The overall result of this increased interest is that we obtained a new and more powerful server this year and are now in discussion with the director of information technology services in the school about the possibility of a .25 FTE student assistant to support repository activities.

This part of the story possibly does not suggest new skills, but I hope that it provides a case study arguing for the notion of a goal of systemic ubiquity for digital preservation. In that context, there is no avoiding the necessity for sophisticated digital skills requiring ongoing maintenance, but our work and its history has shown the importance of getting the maximal number of people in on the act, not by knowing what is good for everyone, but by asking them. There is so much work to go around in this task that we need all the help we can get.

Applied Technical Skills

- Understanding of networking, operating systems, and related file systems, including tools to work within such spaces and to visualize virtual "original orders"
- Beyond this, understanding of standard system management practices, including security, backup, and software installation issues
- Excellent skills as a user of the desktop systems and web-authoring tools used to generate most of the file types at issue
- In many cases, at least modest digitization skills (we had to ingest some paper documents for efficiency of use)
- Good database skills
- Understanding of message digest tools for establishing fixity data
- Familiarity with text analysis tools for deriving document-level access points
- Familiarity with the metadata mining process and tools, including forensic tools and various kinds of metadata and file format registries
- For work with websites, familiarity with web crawlers and how they work

6. *George Washington University's Special Collection's Transformation into a Repository with Digital Services*

Jennifer King, George Washington University

This case study discusses the new equipment, new skills, and new training program established by Special Collections at George Washington University in an effort to meet the needs of our increasingly technologically sophisticated users and to capitalize on the multiple use capability of digital materials. This challenge is one currently faced by many institutions in this Digital Era namely how to best serve our users and at the same time retain our position as appropriate custodians for the ever increasing amount of digital objects in our possession.

Before the Digital Era

In the winter of 2005, the Special Collections department of the George Washington University improved our equipment, procedures, and staff skills related to technology and dramatically increased our ability to create and manage digital materials. Prior to this, Special Collections neither produced nor managed digital materials on a significant scale. We did not have a departmental scanner, our local server quickly became overloaded if used to store digital content, and we had limited delivery options. This situation became untenable because of changes within the archives community as we encountered researchers requesting digital content and acquired donations arriving in a variety of digital formats. Our patrons no longer ask if we can produce a digital surrogate or if we have materials available online. Rather, they expect to receive a digital copy via e-mail or the URL for a collection's digital content. The department realized that if we did not make immediate changes, we would be unable to provide researchers access to historic materials.

The Digital Era begins

Two major changes moved the department into the digital era. First, our new department head initiated the library's purchase of Re:discovery to serve as our collection management system and RFI, Re:discovery for the Internet, to serve as the primary access tool. Re:discovery replaced a combination of Microsoft Office products including Access databases, Excel spreadsheets, Word documents, and paper documents that had been used to manage the department's three collections, Washingtoniana, University Archives, and the Kiev Judaica Collection. Re:discovery allows use of the same data to produce different types of output. From each Re:discovery record we have the capability to produce EAD finding aids and metadata for digital objects for our online catalog and Greenstone databases. We now have a system to manage and disseminate digital materials.

The second change occurred, as part of George Washington University's membership in the Washington Research Library Consortium (WRLC). One of the missions of the WRLC is to provide consortium members with technologies to support digital collections. Toward this end, the WRLC operates two instances of DSpace, the Research Commons meant for use by faculty members, and the Digital Object Catalog (DOC) which is meant for the use of special collections. Using this institutional repository solved our potential storage issues. Now we had the ability to manage, store, and provide digital materials. With these two infrastructure needs met

the staff could focus on how we planned to function within this digital world. Three questions needed to be addressed: What equipment would be required?, What skills would be needed?, and How would we train staff?

Equipment

The Special Collections department acquired some free products and purchased other hardware and software based on our assessment of current and future needs. The table below lists these products.

Technology equipment acquired

Hardware	**software**
Fujitsu fi-4220C (sheet feeding)	PixEdit
Epson Expression 1000 XL (flatbed)	Photoshop Elements
Minolta PS 7000 (overhead)	Adobe Acrobat
DVD burner	Dreamweaver
	Textpad
	Re:discovery (collection mgt.system)

Training

The first critical component of training was the identification of the new skills the staff needed. We identified basic skills by looking at the work we were currently doing. For example, all staff members needed to know how to scan and clean up digital surrogates. They also needed to know how to use OCR software and to search and edit records in the Re:discovery database. For this basic skill set we wrote and provided to each staff member step-by-step procedures and hands-on training.

Advanced skills needed by selected department members included creation of derivatives and EAD finding aids and navigation in DSpace. Individual staff members were assigned to master each task and provide guidelines and tutoring for his or her colleagues. For example, one staff member led the container list conversion project. She prepared guidelines and taught the necessary skills to staff members working with her. Because of her excellent written tools and training, the project continued seamlessly when she left the department. Our work in creating resident experts, training manuals, and departmental tech support continues. We are aided with this technical support by the library's IT department and the staff at WRLC.

See the skills summary for a more complete list of new skills for the digital era.

Evaluation of our progress so far

Most staff members have mastered the Re:discovery and scanning skills necessary to complete current projects and assist researchers. Document and image editing skills have proved harder to teach because it is impossible in our step-by-step guidelines to anticipate all the issues a staff

member will encounter. Therefore, we have come to realize the importance of our mentoring procedures as a training tool.

With our new digital capabilities comes the need to communicate with donors and researchers. For example, we need to be able to discuss options such as file format and size and delivery methods with researchers. This requires understanding the technology and a fluency in IT-ese. With donors of paper records we need to explain how new technologies will increase exposure, access, and usage of the donation. We need to convince donors of born-digital materials that we have the technical expertise and archival training to care for this material.

George Washington University's Special Collections department is still a work in progress. Fortunately, we have one of the most important pieces as we move forward, that is a culture that embraces digital technologies and accepts the responsibility to master the required technical skills. In retrospect I believe this was accomplished in two ways. First, no one was expected to master every skill immediately; rather, each staff member focused on a manageable part. Second, once staff had mastered the assigned skill, he or she became the resident expert. This sense of ownership enhanced this person's identification with this new technology. Of course, new technologies appear frequently and new projects emerge from our interactions with constituents. We need to remain eager to learn and include them in our daily work.

Skills Summary

The technical skills identified as essential for work in our current digital era at George Washington University encompassed three broad areas: digital surrogate creation and manipulation, data transfer for retrospective conversion projects involving the import and export capabilities of a variety of systems, and access tool creation and enhancement. The commitment to this digital era and the work it will involve will necessitate additional technical skills not yet mastered or even understood.

Basic skills

Scan reflective images, film images, and transparencies
Document and image manipulation
Text manipulation using a text editor
Master edit and search functions in Re:discovery

Advanced skills

Create EAD finding aid using XML schema
Web page design
Knowledge of XML, HTML, and SGML
Encode files using Dublin Core, METS, and TEI
Navigate DSpace: creation of collections and object metadata
Master import and export functions in Re:discovery
Operate file transfer protocols

7. *Text and Bitstreams: Appraisal and Preservation of a Natural Electronic Archive*

Maria Esteva, School of Information, University of Texas at Austin

This case, aimed at appraising and preserving an electronic archive, defines the concepts and the tools and methods needed to approach the study. The attributes characterizing the archive at hand led to the development of the concept of natural electronic archives that would allow transforming the archive into a unit of analysis. Furthermore, the methodology for appraising the archive is also somewhat novel, in that it combines fairly common tools used in the archival discipline, while adding others—such as text mining and social network analysis—taken from other fields. A strategy for preserving the archive combines archival considerations such as transparency of transfer and maintenance of the archive's integrity with systems administration skills. It also includes research on the technologies.

Introduction

This case, aimed at appraising and preserving an electronic archive, offers the opportunity of sharing with you the concepts I define, the tools and methods I use, and the skills needed to approach the study. The attributes characterizing the archive at hand led to the development of the concept of *natural electronic archives* that would allow transforming the archive into a unit of analysis. Furthermore, the methodology for appraising the archive is somewhat novel, in that it combines fairly common tools used in the archival discipline, while adding others—such as text mining and social network analysis—taken from other fields. A strategy for preserving the archive combines archival considerations, such as transparency of transfer and maintenance of the archive's integrity, with systems administration skills. It also includes research on the technologies and circumstances involved in the archive's construction over time to fill technical metadata gaps.

The case study

The study involves the appraisal and preservation of the networked electronic archive of a private multi-national philanthropic foundation whose activities span from the beginning of massive incorporation of computer networks in the workplace until 2005. Beginning in the mid-1980s with a few networked stations with minimal storage capacity and the DOS operating system, the institution moved in the early 1990s to a Windows networked system of servers and PC clones. All staff members were linked to the network from their personal desktop computers.

An abrupt interruption occurred during the transition between DOS and Windows. Except for the data residing in the institution's database systems, individual text files, spreadsheets, and schedule data were not migrated to the new environment and were left behind on removable disks of various formats whose lack of clear labeling makes their contents hard to identify. During the summer of 2004, with the help of the institution's systems administrator, emulation and migration tests were performed and it was possible to retrieve text and spreadsheets with their

original look and feel through the use of the originating programs in the DOS interface available in the Windows 1998 operating system.

The current Windows compatible electronic archive containing records and applications dating from 1992 to 2005 resides on a networked server. Three broad categories form the archive: (i) applications, (ii) records, and (iii) e-mail mailboxes, occupying a total of 8 gigabytes of server space. The reason I group all the elements under the electronic archive is because they all form part of the same technical environment and it is not yet clear whether separating the components can be achieved without rendering them incomprehensible.

Under the applications category I include custom made and proprietary scheduling and library software as well as grant tracking and financial applications with embedded data that were bought, commissioned, and licensed beginning in 1986. Included also are pieces of software used for server administration purposes, printer drivers, back-up executables, and other pieces of software whose use was discontinued. E-mail messages were routinely downloaded and resided on the hard-drives of the individual work stations, but in the final five years a few heavily used e-mail accounts were backed-up to the server. To create and store records including text documents, images, spreadsheets, and presentations, each staff member had a directory with identifying initials on the shared server from which—with few restrictions such as viewing some executives' records—they could access each other's files. This structure enabled collaboration through copying, cutting, pasting, and editing of texts across the network. Within their directory, staff members organized their files in sub-directories according to their own criteria and named their files differently, some by using some kind of formal naming convention, others without regard for length and blank spaces. Pieces of text, drafts, versions, and personal records could be found in many folders of this archive. All staff members managed the contents of their shared folders on the server and the e-mail on their own PCs, deciding what to retain and discard.

As technology evolved, new hardware and software were acquired and new records created. In the active networked environment, migration between Windows compatible software provided a temporary layer that allowed for smooth transitions between old and new. By the time older files became unrenderable, the staff did not use them anymore, had used them as templates to create other records, or created new ones that conformed to the current environment. In time, some applications became non-functional and older files could only be opened with some degree of loss. In this sort of semi-functional archeological site, staff members completed most of their daily activities.

Natural electronic archives

Upon my observation of this archive I concluded that the manner in which records were generated and kept, and applications were maintained or removed, resulted in an archive that could not easily be ascribed to digital archiving models currently discussed in the literature. The latter focus more on the creation of sound electronic records within electronic recordkeeping systems than on the way in which digital archives exist (Bearman & Trant, 1997; Cox, 1997; Duranti et

al., 2002; InterPARES Authenticity Task Force, 2002). This led to developing the theoretical concept of *natural electronic archive* through which I could analyze the case in a systematic fashion.

The concept of a natural electronic archive builds partly on that of "natural collections" proposed by Phillip Cronenwett to describe collections of literary manuscripts as they leave the hand of the creator, in contrast to artificial collections such as those fragmented and sold to different collectors (Cronnenwett, 1984, p.106). I suggest that this concept is relevant both to the case study at hand, as well as to archives of public or private persons and institutions showing similar characteristics. Creation of natural electronic archives involves a set of *ad-hoc* practices developed as people adjust to and learn how to use information technologies. A natural archive is not designed or managed by records managers or archivists. Instead, it is what those working in institutions, in different capacities, using different technologies, and making decisions, make of it.

In a natural archive, records are created, named, destroyed, or retained according to individual work practices. Each record creator decides on structure and naming conventions for files and folders, spontaneously or consistently, according to individual mnemonic rules or the spur of the moment. Within the virtual folders, images, spreadsheets, texts, websites, databases, backups, and applications live together under the same roof, placed or misplaced, in organized or disorganized fashion, with or without descriptive clues. Lacking a catalog or aid to help find things, record creators trust that they will remember what and where things are.

Record creators create, re-invent, and leave behind natural archives or parts of them in iterative fashion. Within these iterations, archives evolve towards more structured forms or become interrupted. During these processes, pieces of the archive are left like discarded artifacts in an archaeological site because there is no time or need to go back to old files that cannot be found or opened. Only the latest iterations of records and systems move forward to what seems to be, or is indeed, more usable and efficient. This constant advance and backtracking in the archive is intensified by emergent technologies and ways of creating and storing records and the appearance of new users. At any time, a formerly new way of constructing the archive may be superseded; its criteria, passwords, naming conventions, and logic left behind. A natural archive is an aggregate—or better—an accretion of trials and errors. The evidence of this resides in the vestiges of directories and files left in storage spaces and in the lack of consistency in naming, organization, versioning, keeping, and discarding that characterizes them. In this sense, exploration of natural archives allows uncovering as many recordkeeping patterns—or non-patterns—as there are members of the organization.

In a context without explicit recordkeeping rules, bits and pieces of text are ubiquitous inhabitants. Either shared by different members of a network or used repeatedly by their creator, they constitute the core of many records. This repetition of fragments afforded by the cut and paste function of the text editor, speaks as much of provenance, group collaboration, and fair use, as of hierarchies and corporate culture. As a consequence of what has been described, records within a natural archive are difficult to identify and lack formal documentation. This creates

doubts about their authenticity, their capacity to provide evidence, and the possibilities for preserving them.

The natural electronic archive concept grew while I was surveying the electronic archive and from the interviews conducted with staff members who had worked in the organization since it opened. The former activity consisted of documenting the structure of the server's directory and subdirectories, identifying file formats and applications (with the help of the institution's systems administrator and web resources), and analyzing the contents of random samples of records contained in the staff's virtual folders. The interviews revolved around records creation and recordkeeping practices, work processes and collaboration, and the technologies and circumstances involved throughout the building of the archive. Thus, relevant skills used in this phase of the study include systematic observation and documentation as well as qualitative interviewing.

Problems to address

For legal reasons, the electronic archive must be kept functional for ten years after the institution's closure. While its final destination has not been determined, the way in which it will be transferred to temporary custodians and maintained during this period is fundamental to its long-term preservation. Tasks to undertake at this point are appraising the archive and planning a preservation strategy for the next ten years, with the perspective that in the future all or parts of it will be permanently retained.

Appraisal

The appraisal method is rooted in concerns expressed by Peter Boticelli in his study of networked organizations (Boticelli, 2000). It considers the need to document dynamics and changes in organizations and it explores the meaning of evidence and archival bond—understood as the "network of relationships between records"—in an ambiguous environment (Duranti & Guercio, 1997). Its main departure from other appraisal methods is that it uses digital tools to analyze a large corpus of records inductively.

Text mining and social network analysis use computing algorithms to discover knowledge about the relations among electronic records. By measuring the similarity between records produced and co-produced by staff members within frameworks of time and provenance, the strength of relationships between records and between the staff members and/or functions that created those records can be established. In turn, by averaging the similarities between records of every other staff member or function across time, organizational structure and functions and correspondent changes in dynamics emerge. To confirm the validity of the findings, results are contrasted against the narratives of staff members detailing whom they collaborated with and in what. In this way, the evidence provided by the electronic records in this natural archive will be attested.

An example, only relevant for illustration purposes, shows the logic of the appraisal process. Figures 1 and 2 show a matrix of the similarities between 12 records belonging to four staff members and its correspondent network map.

Figure 1.

		AC1	AC2	AC3	JEO4	JEO5	JEO6	JXM7	JXM8	JXM9	NHIN10	NHIN11	NHIN12
1	AC1	1.000	0.050	-0.158	-0.120	-0.140	-0.221	0.225	-0.198	-0.057	-0.133	-0.257	-0.052
2	AC2	0.050	1.000	0.036	0.443	0.282	0.121	-0.256	0.035	0.137	0.436	-0.232	0.120
3	AC3	-0.158	0.036	1.000	-0.048	-0.165	-0.143	-0.157	-0.122	0.071	-0.108	-0.235	0.037
4	JEO4	-0.120	0.443	-0.048	1.000	**0.681**	0.300	-0.276	-0.026	0.179	**0.851**	-0.283	0.082
5	JEO5	-0.140	0.282	-0.165	0.681	1.000	0.268	-0.256	-0.026	0.107	**0.648**	-0.239	-0.014
6	JEO6	-0.221	0.121	-0.143	0.300	0.268	1.000	-0.230	0.044	0.062	0.264	-0.133	-0.244
7	JXM7	0.225	-0.256	-0.157	-0.276	-0.256	-0.230	1.000	-0.221	-0.100	-0.310	-0.215	-0.242
8	JXM8	-0.198	0.035	-0.122	-0.026	-0.026	0.044	-0.221	1.000	-0.092	0.037	-0.070	-0.231
9	JXM9	-0.057	0.137	0.071	0.179	0.107	0.062	-0.100	-0.092	1.000	0.137	-0.207	-0.218
10	NHIN10	-0.133	0.436	-0.108	0.851	0.648	0.264	-0.310	0.037	0.137	1.000	-0.248	0.063
11	NHIN11	-0.257	-0.232	-0.235	-0.283	-0.239	-0.133	-0.215	-0.070	-0.207	-0.248	1.000	-0.233
12	NHIN12	-0.052	0.120	0.037	0.082	-0.014	-0.244	-0.242	-0.231	-0.218	0.063	-0.233	1.000

The closest records in the map (JEO4, JEO5, and NHIN10) are all board progress reports of different months within the same year. Records in the margins of the network are of personal nature or relate to functions other than the Board's.

Figure 2.

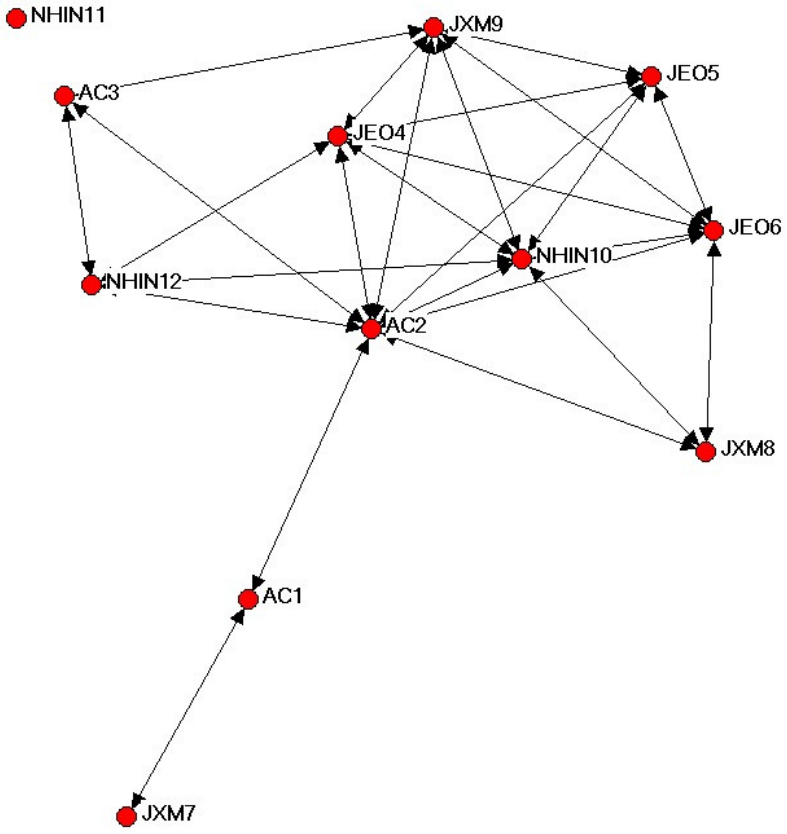


Figure 3 shows a matrix of relationships between staff members based on the similarity between the 12 records analyzed above. The strongest relationship between different staff members corresponds to JEO and NHIN.

Figure 3. Strength of relationships between four staff members

	AC	JEO	JXM	NHIN
AC	0.170	0.182	0.145	0.137
JEO	0.182	0.360	0.148	**0.215**
JXM	0.145	0.148	0.084	0.076
NHIN	0.137	0.215	0.076	0.082

Analyzing the content of strongly and poorly related records will tell us what characterizes relationships between records—provenance, date, type of record, contents—and whether these features can be mapped onto conceptualizations of archival bond. It will also explain the role of drafts, versions, and non-records by finding the proportion in which they exist in the natural archive and how close or not they are from complete records. For the appraisal I will use copies of the records located in the networked archive. The original archive will be the guarantee of provenance and original order.

To carry out this appraisal method, an array of digital tools and research skills are needed. Text transformation and file management software are indispensable to prepare records for analysis. Basic knowledge of UNIX commands is required to run text mining software on a server, and an understanding of quantitative research design and methods is needed to understand the statistical calculations performed by the social analysis software. Considering the volume of records at stake, issues of storage capacity and memory available for processing constantly come up. There is no single piece of software that performs all the steps involved in this appraisal method. Ideally, programming skills would have allowed me to create and manipulate software to improve calculations and analysis. Instead, I had to learn how to use ready-made software and rely on engineering students to do the coding.

Preservation

As described, an electronic natural archive is conceived as a conjunction of processes, tools, and records. In this context, provenance relates as much to network administration and how the different file systems maintain the relationships between records and directories over time as it does to authorship and collaboration. At this point, when the archive as a whole must be retained for legal accountability, a decision was made to preserve the archive intact, within the same structure as it was used but in a newer (though compatible) technical environment. This will allow testing an archival protocol for transfer and maintenance and exploring the archive's technological dependencies to devise a preservation strategy. The transfer and maintenance protocol, which will be carried out by an IT consulting company, contains the following elements:

- Copying the contents of the archive in their original structure to a new server (hardware and software) and to other (redundant) storage media upon successful completion of a transfer test trial
- Creating a dark archive on the secure server with a designated PC for access
- Carrying out transfer audits, including pre- and post-inventories
- Technological preservation of the old networked server until mechanical failure occurs
- Bitstream preservation
- Guaranteeing security and maintenance of the new server
- Continuous monitoring of file integrity and renderability
- Documentation of transfer and maintenance processes

Devising the protocol required understanding the technologies (past and present) in the archive such as file formats and their corresponding rendering software, operating systems, and hardware. It was also important to become acquainted with file transfer methods, refreshing strategies, server maintenance, and server integrity checking routines. All of that had to be adapted to archival practices in order to ensure transparency in the chain of custody and to make sure that future events will be documented. While much of this was explored in the technical literature, direct observations of the archive and talks with systems administrators and IT consultants were fundamental.

Upon the transfer of the archive and considering the appraisal results, the next steps and skills that will be explored to devise a preservation strategy are:
- Construction of a technical metadata timeline by researching the technologies involved in the archive since it was created and through automatic metadata extraction from samples of files belonging to key periods of technological changes
- Testing preservation methods of migration and normalization on sample files belonging to the same key periods

Conclusion

This case study allows the exploration of current digital appraisal and preservation methods for a natural archive. It involved acquiring a considerable number of IT skills and combining them with traditional archival practices. To face the challenges of digital archiving, archivists have to become IT specialists and make use of the potential of digital tools to explore electronic records. Currently, text and data analysis are becoming the main tools which researchers in different academic fields use to explore corpora of texts and databases. Their use for appraisal purposes provides the opportunity for archivists to make of appraisal a research endeavor that promises to allow us to explore/define/determine the meaning of archival bond and evidence in the realm of electronic archives. Upon completion of the transfer and the appraisal, in combination with the metadata and migration findings, it will be possible to determine what is at stake in the natural archive, what can be preserved, and how.

Bibliography

Authenticity Task Force. Requirements for assessing and maintaining the authenticity of electronic records. *InterPARES* (2002). Retrieved October 6, 2005, from http://www.InterPARES.org/display_file.cfm?doc=ip1_authenticity_requirements.pdf

Bearman, D., and J. Trant. Electronic records research working meeting, May 28-30, 1997: A report from the archives community. *D-Lib* (1997). Retrieved October 3, 2005, from Archives and Museums Informatics website: http://www.dlib.org/dlib/july97/07bearman.html

Boticelli, P. "Records Appraisal in Network Organizations." *Archivaria* 49 (Spring 2000): 161-191.

Cox, R. J. "Electronic Systems and Records Management in the Information Age: An Introduction ." *ASIS* 23, no. 5 (June/July 1997). Retrieved October 3, 2005, from http://www.asis.org/Bulletin/Jun-97/cox.html

Cronenwett, P. L. "Appraisal of Literary Manuscripts." Chapter 5 in Nancy E. Peace, *Archival Choices: Managing the Historical Record in an Age of Abundance*. Lexington, Mass.: Lexington Books, 1984. Duranti, L. & Guercio, M. (1997). Research Issues in Archival Bond. *Electronic Records Meeting, Session I.* Retrieved March 29, 2006, from Archives and Museum Informatics website: http://www.archimuse.com/erecs97/s1-ld-mg.HTM

Skills Summary

The list of skills is organized around the three major steps involved in my case study. For some skills, I identified where I had to work with other specialists to achieve what I needed. In the process, I realized that acquiring these skills would have allowed having better control of the case.

Skills for developing the concept of "natural archives"

- Systematic observation and documentation
  - Understanding of networked systems (access, passwords, security, servers, connectivity, etc.)
  - Inventory of the directory structure and substructure including annotation of file renderability and applications functionality
  - Mapping the directory structure to organizational functions, personnel and time periods
  - Identification of file formats and applications (work with systems' administration)
  - Use of web resources for identification purposes
- Research and testing (file formats, applications)
  - Emulation and migration tests
    - Installation of old software and hardware pieces to retrieve old files (work with systems' administrator)
    - Migrate file formats within their proprietary software environment
    - Migrate file formats outside of their proprietary software environment
    - Understanding/research of the underlying encoding of text files
- Design of recordkeeping interview protocol

        - Records-creation, recordkeeping, functions, collaboration, work-periods
- Interviewing skills

Skills for conducting appraisal

- Text Mining
    - Information retrieval, data mining, natural language processing theories, concepts and utilities
    - File management and text transformation software for text pre-processing
    - Sampling methods
    - Information retrieval / data mining / text mining software
    - Unix commands and environment (many open source packages work in Unix environment)
    - Matrixes (interpretation, software, graphics)
    - Math to understand the logic behind calculations
    - Storage issues for large amounts of files and large files
- Social network analysis
    - Social network analysis theory and concepts
    - Research design and statistics
    - Social analysis software
    - Matrixes (interpretation, software, graphics)
    - Math to understand the logic behind calculations
    - Storage issues for large amounts of files and large files
- Programming (worked with a math graduate student to perform large calculations and better automate text mining processes)
    - Mat Lab
    - C or C++

Skills for digital preservation

- Technological history of the archive at hand
    - Applications
    - File formats
        - Underlying encoding of text formats
    - Operating systems
    - Server technology
    - File systems
- Migration between operating systems and across hardware
- File transfer methods and protocols
- Automatic inventorying of records in Windows or Unix environments
- Systems administration routines and software
    - Security, disaster planning, back-ups
    - File integrity checking
    - Event tracking

    - Standard operating procedures (determine what needs to be reported in this case and write an SOP that will function across time)
    - Programming to develop an automatic file rendering program
- File normalization and transformation software
    - Software installation and configuration
    - UNIX commands/compiling software
- Metadata standards (technical, administrative, structural, descriptive, etc.)
- Metadata extraction/conversion/ tools, viewers, and related utilities
    - XML
    - XSLT (ideal to grab what is needed from the XML document and present it according to standards)
    - XML editors
    - JAVA programming (ideal to modify/configure existing extraction software or to create metadata extraction software)
    - Understanding of file properties and changes of file properties across applications and operating systems
- Design of technical interview protocol
    - IT decisions, technologies, programming languages, platforms, systems administration routines
- Interviewing skills

8. *The Next Generation Finding Aid: The Polar Bear Expedition Digital Collections: A Case Study in Reference and Access to Digital Materials*

Beth Yakel, School of Information, University of Michigan
Polly Reynolds, Bentley Historical Library, University of Michigan

This presentation and paper will address the vision, knowledge, and skills required to reinvent the archival finding aid for the future and the challenges of providing new types of access to digital materials. Our case study focuses on the development of the Polar Bear Expedition Digital Collections (http://polarbears.si.umich.edu), an interactive website featuring digitized materials documenting the history of the American military intervention in northern Russia at the end of World War I. The project, a collaboration between faculty and students at the University of Michigan School of Information and archivists at the Bentley Historical Library, features a website showcasing more than fifty digitized collections of primary sources, including diaries, maps, correspondence, photographs, ephemera, printed materials, oral history interviews, and a motion picture. The site represents the first example of the School of Information's Next Generation Finding Aids Project, an effort aimed at creating innovative approaches to archival content online.

Scenario

The "American Intervention in Northern Russia, 1918-1919," nicknamed the "Polar Bear Expedition," was a U.S. military intervention in northern Russia at the end of World War I. Since many of these soldiers originated from Michigan, the Bentley Historical Library at the University of Michigan, an archives documenting Michigan history, has collected materials related to this event since the 1960s. As a result of this focused collection development, the Bentley has amassed one of the largest and most comprehensive collections on this topic, consisting of over fifty primary source collections as well as numerous published materials. As its first large-scale experiment in digitizing entire archival collections, in 2004, the Bentley Historical Library chose the frequently used Polar Bear materials to digitize in order to increase access and protect and secure the fragile originals.

In 2005, faculty and students from the University of Michigan School of Information (SI) began a research project investigating "Next Generation Finding Aids" with the goal of reimagining traditional finding aid structure and implementation. Current online finding aids often reproduce the structure of a paper finding aid without taking advantage of the electronic environment. The digital realm allows for searching, interlinking, collaboration, and interfaces beyond text—properties that a paper finding aid does not possess. Additionally, repositories have not considered the interactive potential of the web or experimented with incorporating social navigation features into finding aids. *Social navigation* is defined as use of the navigation of others' paths to guide and structure the activities of future users within that space.[66] Furthermore, while many repositories employ Encoded Archival Description (EAD) in their online

[66] P. Dourish and M. Chalmers, *Running out of Space: Models of Information Navigation.* Proceedings of the Conference on Human Computer Interaction (1994).

finding aids, none have yet taken full advantage of all of the properties offered by EAD and XML-based systems. The Finding Aids Next Generation (FANG) research group seeks to expand the capabilities of EAD, make the archival and research experience collaborative and participatory, and fully exploit the advantages of the electronic environment for displaying and connecting users to archival content.

The Polar Bear collections at the Bentley Historical Library of the University of Michigan proved to be an excellent set of experimental collections to frame our ideas. First, the Polar Bear collections have a devoted and interested audience. Researchers request these collections for their historical as well as genealogical content. Therefore, we knew that online collaboration and participation would be possible and valuable. Second, the Bentley had just digitized all of the Polar Bear materials in their collection, thus providing us with an existing data set with opportunities for both experimenting with finding aids and envisioning interfaces for linking digital content to finding aids in a meaningful way. Finally, the Polar Bear collections have always been considered one unit, even though they are made up of over fifty individual collections. These collections, therefore, provided an excellent opportunity for us to experiment with uniting and interrelating physically separate collections intellectually, without destroying provenance.

The Polar Bear Expedition Digital Collections project was a joint effort between two units of the University of Michigan: the School of Information, a graduate program with specializations in library science, archives and records management, human-computer interaction, and information economics, and the Bentley Historical Library, an archives collecting materials documenting the history of the state of Michigan as well as serving as the official archives of the University of Michigan. Project collaborators included archivists at the Bentley Historical Library as well as graduate students and faculty at the University of Michigan School of Information, with backgrounds in computer science, archives, and human-computer interaction.

Building on technologies and ideas from collaborative filtering and "folksonomy" systems,[67] as popularized in such websites as Flickr (www.flickr.com), Amazon (www.amazon.com), Wikipedia (www.wikipedia.com), deli.cio.us (www.deli.cio.us.com), and Everything2 (www.everything2.com), the site employs several innovative new features enhancing both access and reference to the collections. On the backend, we enhanced the EAD finding aid, marking up additional subjects and concepts in the scope and content as well as the biographical sections. We then employed Perl scripting to extract and compile related terms and concepts into an SQL database. Researchers can now easily access collections through subjects, geographic locations, genres, individual names, and military units, allowing them to discover new interrelations and links between these collections. We also added hyperlinks within collection scope and biographical descriptions (similar to Wikipedia links) in order to associate related terms within content as well.

[67] T. Hammond, "Social Bookmarking Tools (I)," *D-LIB Magazine* 11, No. 4 (April 2005). Available online at http//www.dlib.org/dlib/april05/hammond/04hammond.html.

Collaborative filtering features such as link paths serve to increase access to the collections in new ways. The link paths, found at the bottom of every page, function as a type of automatic recommender system, relaying immediate feedback to researchers on how others reached a particular item or collection. We hope that link paths will provide alternate and unexpected interrelations between subjects and collections. Additionally, we incorporated a database of the soldiers who were part of this expedition; this database was created from primary and secondary sources from within these collections, published works, and from in materials provided to us by users.

Commenting and login features have enhanced and enlightened the reference process. Since the site debuted in January 2006, 58 individuals have registered. Of those people, 10 (17%) have included biographical statements. While the site allows users to comment on collections and individual items as well as search others' comments, the majority of the commenting has been done in the context of the biographical statement at log in. Already visitors have used this biographical section to discuss their backgrounds and interest in this topic, provide new information about soldiers and the collections, and add links to their own Polar Bear-related web pages.

Users have employed the commenting feature as a method of interacting with the archivists and project staff. At any point, visitors to the site can contact and asynchronously dialog with the archivist. We originally anticipated archivists and reference archivists using the comments to answer user questions or offer research advice and help. Numerous visitors have already identified inconsistencies in the information and submitted new information. As a result we have had to develop procedures for making changes to the site as a result of visitor information. In addition to questions and comments to the archivist, visitors have also offered us collections to digitize and include on the site, an unexpected development. We have forwarded offers to the Bentley Historical Library, resulting in several new additions. The comments and user biographies have contributed to the development of a research community and serve to capture and preserve the knowledge of archivists and researchers, something not possible in a traditional archival setting.

Thus far, we have found most of the dialog to be between "The Archivist," an omnipresent persona on the site, and the users. In one instance, however, a knowledgeable researcher did suggest additional sites and resources to another researcher attempting to find additional information on her ancestor. We hope to encourage and see more user-to-user interaction in the future.

Functionally, two additional features will be implemented in the future. First, a "virtual call slip" generator will aid researchers in requesting collections at the physical archive and may also serve as a means of more easily compiling collection use statistics. Second, an implementation of the increasingly popular "tagging" feature will let users assign their own short descriptions to content. Such "folksonomy" classification systems (as seen in sites like deli.cio.us and Flickr) have enhanced the search and retrieval process as they allow users to im-

plement their own natural language vocabulary and not be constrained by authoritative cataloging terminology. We anticipate that this tagging feature will better support the social navigation within the site.

Technology

The technological backend for this project consists of a 400 gigabyte XServe G5 2.3Ghrz server purchased expressly for our research. Project personnel maintain the server, although it resides within the School of Information's information technology department. We are running a MAC OS X operating system version 10.3 (Panther), but will be moving to 10.4 (Tiger) over the course of the summer.

The Polar Bear Expedition site was implemented using a combination of open source software, including a Perl content engine (Everything2, www.everything2.com), MySQL, and Apache. Open source software not only proved to be more cost effective, but Perl and MySQL systems are also widely accepted and implemented online delivery systems, thus documentation and resources exist for troubleshooting. We also chose open source software so that other repositories would be able to implement our system easily and cost effectively in the future. Additionally, we chose Everything2 as a content management system because it supports many of the social navigation features we wanted to implement. Cascading style sheets form the backbone of the interface design.

The images comprise 80 gigabytes of data. The display images are JPEGs derived from 600 dpi TIFF masters amounting to over 400 gigabytes of data. Images are displayed at 25% of their actual size. All images also have a zoom feature. Images do appear to be different sizes, however, because of some inconsistencies in the actual imaging process. Images were delivered with minimal metadata, creating substantial work on our end to match existing metadata in the EAD finding aid to the images. To date, this has only been partially successful. Due to inconsistencies in the component tag <C0> levels in EAD, we have been unable to write a script, and therefore hand coding has been necessary to link metadata to actual images.

We considered early on whether to use EAD as the basis of the project or to develop an alternate system for delivering archival content online. While we recognize that EAD does have limitations, we ultimately chose to adhere to this standard as it is widely accepted and implemented in archival circles and also will allow other collections to be imported into our system without extra effort. One of the challenges of utilizing EAD, therefore, became dealing with its flexibility. This entailed writing Perl scripts to make global changes and using sheer archivist power to re-code the EAD inventory. Balancing human intervention and trying to optimize machine processing became a key managerial decision.

Throughout the course of our work, we were confronted with a number of challenges that informed the type of skills and knowledge necessary for completing this type of project. Ultimately we discovered that delivering archival content online requires skills that go beyond

simple html and archival principles. Through this project we found principles in computer science, human-computer interaction and library science disciplines to be crucial.

For example, the extended markup of EAD required a basic knowledge of cataloging and Machine Readable Cataloging (MARC) in order to assign additional subject headings. As we discovered, the flexibility of EAD's structure allows for ease-of-use and customization but also leaves much room for vocabulary discrepancies. Electronic environment systems cannot yet make natural language distinctions, thus, information must be consistent in order to be meaningful and useful. Therefore, we spent a great deal of time checking for authorities and standardizing headings. Furthermore, since we combined data from several disassociated electronic sources—a database of all the soldiers, the EAD finding aids, and the MARC records—inconsistencies among these data also required much standardization.

Due to the small size of the Polar Bear collections, the Bentley only originally provided access to many of these collections through MARC records and not finding aids. These MARC records were part of the University of Michigan's online integrated library system (ILS), MIRLYN. As a result we realized that reuse of the subject headings was problematic. For example, all of the collections contained general subject headings such as "Polar Bear Expedition" and "World War, 1914-1918." In the MIRLYN catalog with over 5 million MARC records, these subject headings provided essential information to differentiate cataloged items and assist researchers in honing in on the collections. In our smaller system, however, these subject headings were not specific enough to be useful. This vocabulary discovery has certainly informed current archival descriptive practice.

Knowledge of principles of human-computer interaction (HCI) and usability were also essential. In our background research of archival content delivery, we have found that many archivists do not consider their audience: Who will be using the site? How will they use it? Are these technologically savvy users? How comfortable is this audience with archives? For example, many sites employ archival jargon (like linear feet and finding aids) without any explanation. While such terms may be second nature to archivists, such terms are not familiar to the majority of users. To tackle this issue, we compiled our own archival terms dictionary and hyperlink terms from the content.

Archival content systems also overlook basic principles in design and accessibility. Is information easy to find? Are Gestalt principles being followed? Can you find your way back easily? Does the organization of the site make sense to the user? Through much trial and error as well as ongoing user testing, we aimed to make our site usable, understandable, navigable, and accessible. Despite the complex backend of the website, we wanted to make the overall design simple and clean. We found insight and comments from users, Bentley archivists, and HCI students to be extremely valuable. Through these comments and suggestions we learned about many challenges that users face in using archives, both in a digital as well as the physical environment. We discovered that even the simplest changes, like font size or altering the color, has an enormous impact on how users interact with the site. We followed up on the usability of the

site after launch with a questionnaire and selected interviews with users. Overall, visitors like the architecture geared toward browsing and find the search function easy to use. Visitors are less familiar with the link paths and the bookmarking functionalities. As a result, we have decided to change the name of the link paths to reflect the more familiar metaphor exemplified in sites such as Amazon.com to say something like, "Visitors who viewed this information also viewed…."

In addition to the purely technological knowledge from HCI, we have also learned to heed knowledge and understand the skills needed to foster social navigation and be aware of the social side of computing technologies. In this way we are very conscious of not only our presentation of the content but also the social forum we have created and hope to foster on the Polar Bear Expedition Digital Collections site.

Finally, we cannot discount the necessity of computer technical skills and knowledge. We were fortunate enough to have students with a background in computer science on our team. A basic background in database design, HTML and CSS, scripting, and network infrastructure did allow non-technical project participants to work and effectively communicate with the individuals involved in implementing the back end technology.

Despite the increasing need for newer skills to complete this project, including advanced technical skills, library science basics, and an understanding of human-computer interaction principles, we ultimately found that the basic principles of archival theory continue to apply and in fact, become increasingly crucial as archivists continue to move into the digital realm. We do confirm, therefore, the hypothesis that "what we do remains the same, how we do it changes."

Before this project began, we considered the impact that completely digitizing an archival collection might have. For example, does reading a soldier's diary online change the emotional impact of the content? Does the lack of an intermediary, such as a reference archivist, change how these collections are viewed and used? Such questions remained at the forefront as we considered how to design and deploy this site.

We found that faithfulness to basic archival principles was crucial to influencing how content would be understood and used online. For example, we could have easily combined all of the individual collections into one Polar Bear collection with the individual images accessible via subject or other schemes. However, we decided to adhere to the principle of provenance and maintain the individual soldiers' collections. Thus, while a user is able to access collections via subject, geographic location, and other access points, the user will always be led into the scope/content and biographical content of an individual collection before accessing the digital content. In this way, the user is provided a context for the content (which influences understanding) and also mimics the way that a user would approach the collection in a physical archive.

Furthermore, we adhered to the original order of materials as well, also serving to replicate for users the experience of viewing the physical collections. We did create more intellectual access to the collections by creating virtual folder galleries based on natural divisions of the materials that were not present in their current physical form. As a result, users can see thumbnail images of an entire folder all at once. Folder and content order online is maintained and each page includes a "next arrow" so that users can digitally thumb through collections.

Through this project, we have also found that the reference archivist role has not diminished, but changed. Even though content has been moved online and users no longer visit the physical archive, researchers still require and request the services of archivists. We have built this notion into our system through our commenting feature, and already many of our users have interacted with archivists through this new medium of communication. Thus, reference archivists need to be prepared for alternate forms of intermediary interaction outside of the traditional reference interview. Frequently asked questions, help pages, and online tutorials are other tools that will confront reference archivists in the future.

Specific Skills Needed for this Project

- Advanced technical skills and knowledge of computer programming languages and systems: Perl, Javascript, XML, SQL, HTML, CSSx
- Knowledge of relational database design and use
- Understanding of networks and Internet infrastructure
- Experience in hardware and software use, including security and maintenance
- Principles of human-computer interaction and usability
- Familiarity and experience with cataloging principles including MARC, Anglo-American Cataloging Rules, and authority control
- Archival principles and languages, including an understanding of EAD, XML-based systems and stylesheets
- Consideration of ethics of Internet use: copyright, privacy, and security
- Knowledge of social navigation and designing systems to facilitate and support these mechanisms
- Current trends in Internet access and use: tags, blogs, comments, recommender systems, wikis, and other types of "folksonomy" systems

9. *Archival Reference Services for Digital Records: Three and a Half Years Experience with the Access to Archival Databases (AAD) Resource*

Margaret Adams, Electronic and Special Media Records Services Division, National Archives and Records Administration

Set within a discussion of NARA's custodial program for electronic records and the reasons for the development of the Access to Archival Databases (AAD) tool, this case study explores the impact of AAD on NARA's reference services for electronic records. Has the availability of AAD changed its nature or the nature of reference services in the traditional still picture or textual records units? Has there been a change in the research community served by NARA's electronic records program or in the types of services expected by the public? In the course of this scenario explication, the case study implicitly considers the evolution in the skills archivists have needed to offer reference services for electronic records. We include a discussion of some "generic" lessons learned from NARA's reference experiences related to digital records generally, and through AAD in particular. In conclusion, skills are discussed briefly in the context of the digital environment.

Scenario: NARA's Custodial Electronic Records Program

The U.S. National Archives and Records Administration (NARA) has the oldest and largest custodial program for archival electronic or digital records of any traditional archives.[68] By "traditional" archives we mean archival institutions whose foundational holdings are analog records. NARA's custodial electronic records program has been part of various NARA organizations over the years, and currently is a division of the Modern Records Program, Office of Records Services, Washington, D.C. The holdings of NARA's custodial electronic records division are born-digital federal records and measure in the terabytes. The division is not currently engaged in any major digitization projects.

Processing electronic records into the National Archives, preserving and maintaining them over time, and providing access to them has always involved working with computer hardware, systems, and other software. Thus since its beginning, the professional archival staffs of the electronic records custodial program have had varying combinations of archival and computer-using skills, along with additional subject matter expertise. They have worked in partnership with technical colleagues within the program whose education, training and/or experience, as well as job responsibilities, were as computer programmers, information technology specialists, or systems analysts. This has been true regardless of the size of the staff, and whether additional contract staff have been involved in specific projects.

Over the years staff have possessed a wide variety of education and experience. Some have had careers as archivists of analog records. They became digitally savvy as their work in the electronic records program provided an opportunity for learning not only how and when to use computers, but why and how federal agencies were employing computing technology for records creation, compilation, or use. Others came to NARA with interdisciplinary backgrounds,

[68] For the history of NARA's electronic records program see Bruce I. Ambacher, ed. *Thirty Years of Electronic Records* (Lanham, MD: Scarecrow Press, 2003).

including history or another of the social sciences, along with training, self-education, or interest that gave them confidence or at least a degree of comfort using the computer, sometimes including programming. They learned the practice of archivy on-the-job, with specialized training as needed. As computing capabilities evolved over the decades, so did the staff's understanding and use of technology. More recently some staff have come to the program from archival education programs; since personal computing became ubiquitous in the 1990s, all new professional hires, regardless of their prior education and experience, have been expected to be facile in using a personal computer and a variety of software applications.

Reference Demand for Electronic Records at NARA

To date, numeric and alpha-numeric data files, the first type of digital archival records, a by-product of data processing, constitute the largest percentage of NARA's accessioned electronic records series. Electronic records in structured data files generally are not narrative in nature and rarely are meaningfully eye-readable, even with software that supports display of the raw data. Because they are not visually comprehensible, nor physically tangible, they possess none of the affordances of analog archival records. Nonetheless this type of primary source is the basis for much of the social scientific research undertaken since at least the second half of the twentieth century.

Until the release of the Access to Archival Databases (AAD) tool in February 2003, NARA's basic reference services for electronic records involved offering information *about* the records, limited service providing information *from* the records, and for a cost-recovery fee, providing copies of digital data files and their related technical information (documentation). Most documentation is analog and on paper. Copies of the preserved digital data files have been reproduced on a variety of computer-readable media, changing as new media have emerged in the marketplace, and encoded to meet the user's specifications. Further context for this type of reference services provides perspective. As recently as a decade and a half ago, the community "users of archival electronic records" generally was limited to persons who were applying some form of social science research methodology in their work. Few in this community use NARA's analog records in their research, nor do they tend to visit NARA in person to consult finding aids or archivists. As a result, electronic records reference archivists at the National Archives long ago became adept at remote reference services, adapting the intended expectations of reference interviews and subsequent intermediary reference services to the realities of distance communication. Formerly this was by telephone or postal mail but now is almost exclusively via e-mail. Staff expend considerable effort preparing online descriptive resources, and the requests received indicate that these materials are being read.

For the remote archival researchers mentioned above, electronic records in structured data files are primary sources of choice. Their expectations have been that NARA would identify, acquire, process and preserve valuable federal digital data files so that anyone subsequently could find or receive information about them and obtain copies of the files. They then would be free to analyze the records on their own terms and with whatever computing hardware and soft-

ware they had or had access to, retaining copies of the records indefinitely. For this researcher community NARA's "traditional" form of access to electronic records remains highly desirable, although now researchers probably would prefer to download files from a NARA server to theirs, rather than ordering files on removable media. But the interest in acquiring their own copies of the files remains.

Over the thirty plus years of NARA's custodial program for electronic records, experience with reference services has led us to generalize that the universe of potential users of digital records currently in archival custody likely falls into one of two groups. One consists of persons involved in research that is intended to create new knowledge or understanding, a category that defines much of the archival research of academics and other "professional" researchers. Among this group are the researchers described earlier for whom NARA's "traditional" form of electronic records access was originally developed and continues to be well suited. The other general category consists of persons seeking archival materials as a source of specific factual or personal documentation; they seek individual record-level access rather than copies of files or other aggregates of digital data. Some forms of historical research are of this type. We tend to refer to this group as "information-seekers" and in the contemporary digital age, as in the past, their numbers far exceed the former group of researchers. The two groups and their expectations of public archives are usually distinct, though often not mutually exclusive. All of this is similar to the universe of potential users of analog archives.[69] What is unique to the digital world, however, is that an infinite number of people, whether selecting archival records to test hypotheses or information seeking, can simultaneously use the exact same digital record(s), and without being bound to any particular place or time.

Even before the advent of personal computing and the related expansion in the types of digital records that federal agencies create using the relatively newer information processing tools of personal computers, and well before the emergence of the World Wide Web, there was a growing population seeking information from NARA's electronic records who were not well served by the "traditional" form of access to them. For many, the fact that the information they sought was preserved in a digital mode was not relevant; they were seeking documentary information, wherever and however it might be found. A well-known example illustrates this point. Following the early 1980s transfer into the National Archives of electronic records of casualties of the Korean and Vietnam wars, veterans seeking information about their buddies or to document their claims for post-traumatic stress disorder, civic groups planning to honor casualties from their communities, and families seeking information about the deceased or missing, came to expect that NARA staff would, in response to their requests, retrieve this information from the casualty data files. This focused demand presaged more broadly based expectations for record-level access once desktop personal computing and especially the Internet became

[69] These generalizations are similar to, among others, those of George Chalou in "Reference," *A Modern Archives Reader,* eds., Maygene F. Daniels and Timothy Walch (Washington, D.C.: National Archives and Records Service, 1984), 48-50; Mary Jo Pugh, *Providing Reference Services for Archives & Manuscripts* (Chicago: Society of American Archivists, 2005), 33-73; and Ann S. Gray and Diane Geraci, "Complex Reference Services: Data Files for Social Research," *The Reference Librarian* 48 (1995): 135-137.

ubiquitous, and the "E-FOIA" amendments to the Freedom of Information Act became law. As mentioned above, the fact that the information sought was preserved in electronic records was an irrelevant consideration, except when that reality seemed to create a barrier to access. NARA's electronic records reference staff responded to much of the demand for individual casualty records by using statistical analysis software or a variety of customized computer applications, including some that had been developed for archival processing purposes. Nonetheless, these techniques offered only limited possibilities for attempting to meet Internet-era expectations for record-level access to digital records on many subjects. Meeting rising access expectations clearly required developing an automated generic access tool for fielded data, and by the late 1990s, the technical capabilities became available for doing just that.

Access to Archival Databases (AAD)

After several years of planning and development, and building upon all prior experiences processing and offering reference services for electronic records, in February 2003 the National Archives launched AAD, the Access to Archival Databases (www.archives.gov/aad). This resource was developed under the auspices of NARA's ERA program, with a contract to SAIC, and with significant involvement of the staff in the custodial division with responsibility for electronic records, and others.[70] In time, staff from NARA's Still Picture unit and the Civilian Textual Reference unit also became involved. When first released, AAD was an ORACLE-based application, featuring online access to a selection of 32 series of born-digital fielded data from accessioned electronic records. The series in the initial rollout contained approximately 50 million records, selected because records in their files identify specific persons, geographic areas, events, transactions, organizations, or index records in NARA's analog holdings. New series have been added to AAD continuously, and at this writing AAD provides the potential for record-level retrieval from a highly heterogeneous mix of more than 86 million records in 46 series.

AAD provides series and file-level descriptions for the records it contains, offers search and retrieval capabilities for online access to its selection of NARA's "most in demand" digital records with the metadata needed for understanding the meaning of the records, and for some series or files also includes additional technical information scanned into PDF documents. Essentially, AAD has introduced a limited form of self-service reference. The number of users who now independently search for, retrieve, print, or download records online is significantly larger than the number who previously requested information about or from NARA's accessioned electronic records. In September 2005, AAD introduced a new capability and for the first time provided direct access to individual digital photographs of disasters from the Federal Emergency Management Administration (FEMA), as well as to the index records that identify them.

---

[70] As mentioned, AAD was developed with the support of NARA's Electronic Records Archives (ERA) program, and is the first publicly accessible application developed under the auspices of the ERA program. At the present time, however, the division with custodial responsibilities for NARA's electronic records is organizationally separate and distinct from the ERA program.

Indexes to two series of still picture records from NASA (National Aeronautics and Space Administration) had been added to AAD at an earlier point. In December 2005, in direct response to two sets of usability studies and feedback from a voluntary online customer satisfaction survey, NARA rolled out a redesigned AAD. AAD's interface was modified to take advantage of innovations in online searching technology that occurred after the original AAD was designed. The redesign supports free-text searching, while continuing to offer fielded searching, and combined free-text and fielded searching. Search and retrieval became more contemporarily user-friendly and free-text searching of digital text records became efficiently feasible. In March 2006, the first series of digital narrative records, telegrams from the Department of State's Central Foreign Policy Files, 1973 and 1974, were made available online through AAD.

AAD is publicly accessible via the Internet, 24 hours a day, seven days a week. Use measured by queries has grown steadily since the first release of AAD, spiking each time there is some kind of media focus on the resource. By spring 2006, queries for records averaged in the neighborhood of 5,000 per day. Queries run for records in the newly added digital series from the State Department have increased rapidly, making it one of the more popular AAD series, and apparently attesting to AAD's success in reaching at least a portion of the traditional archival research community.

Most of the digital records selected for AAD were not created for any public information purpose and so readying them for online public access is a significant and labor-intensive undertaking, customized series-by-series as necessary. The records were created by federal agencies over many decades to meet their own programmatic or business needs, using whatever generation and type of technology they had. As a result, historic digital records often contain what to a lay person may seem to be inexplicable idiosyncrasies, as well as an almost infinite variety of content issues. The metadata preparation needed to explain all of this is complex and extensive, well beyond the archival description written for all accessioned records series. As a starting point, AAD's metadata comes from the metadata originally prepared by the electronic records staff responsible for verifying the records as part of their accessioning, and is subsequently enhanced so that every field and every code table is included. Contract project staff have handled the initial AAD metadata preparation phase and make copies of the preserved data files to be included in AAD, while archival staff review the metadata, write additional explanatory notes at the field level, as needed, and test functionality and public readiness for each new file.

Impact of AAD on Reference Services

Offering even limited self-serve access to electronic records seems to be altering the balance between the volume of routine and "information from" reference requests and the volume of more complex requests. Overall, interpersonal demand ebbs and flows, as for all reference programs; during the current year it has been rising. Since AAD the proportion of routine and "information from" inquiries has declined while a slightly larger proportion of all requests now are research-oriented. Our colleagues in the Still Picture and Textual Reference units tell us that the most evident impact of AAD on their reference demand is that they now can refer research-

ers directly to AAD to do their own searching for records, when requests are for information about or from records that are accessible through AAD. This is similar to the experience of the electronic records reference staff, who, since the first months of AAD, have used a referral to AAD in over 35 percent of the replies to reference requests. But, as described above, facilitating and simplifying some users' access to electronic records in the manner of AAD has raised the ante for the archival staff. Employing this type of technological solution for handling the most routine, common, or straightforward requests for digital data records or information from them requires extensive preparation of the records and their metadata.

In terms of expectations related to record-level access, reference staff now field a larger number of questions from researchers perplexed by the records or the information in the records they have found, with the "finding" likely only because of the ready online accessibility of the records. New variations on basic remote communication skills are thus required. Among the more surprising requests that have become frequent since online accessibility to electronic records have been those from people seeking "correction" of the information they have found in archival records, a drastically different version of the challenge of maintaining authenticity of records in a digital era than the challenges theoreticians have posed. Reference archivists need to hone their persuasive skills as they decline to "correct" records, explaining the archival preservation mission to a lay public in the process.

Queries in an online archival resource and personal responses by an archives' reference staff to requests directly received are not comparable. It nonetheless is worth noting that the current average of daily AAD queries (5,000) exceeds the annual level of inter-personal requests for information about or from records ever experienced by NARA's custodial electronic records program. Consistently, the vast majority of the queries run in AAD are for records in series that identify individual persons; a majority of more than 60 percent of the voluntary respondents to the online AAD customer satisfaction survey routinely self-identify as genealogists or personal historians.

Overall, there seems to be enhanced awareness of NARA's electronic records custodial program as a result of the presence of AAD on the Internet. Remarkably, and gratefully, the high volume of querying in AAD has not caused a significant increase in direct requests received by NARA's reference services staff responsible for electronic records. Brief exceptions have occurred during the occasional rare periods of AAD malfunctioning, or a storm-induced off-line period. About 20 percent of the requests received by the electronic records custodial reference program in the first half of fiscal year 2006 related in some way to AAD. The majority sought additional content related to records "found" through searching in AAD, suggesting that "finds" spurred further interest. But 20 percent of the approximately 1,000 direct requests in this period is an infinitesimal measure compared to the volume of AAD queries (over 800,000) run during the same time. We attribute this outcome to the thorough and dedicated preparation, review, and testing of the records and their metadata that precedes the availability of any series in public AAD, as discussed above. In addition, the series included in AAD account for most of the series

of digital records for which, in the past, NARA regularly received heaviest demand for record-level access, or "information from the records."

Direct reference requests received in the electronic records program have increased a bit in recent months, but the nature of many of them suggests that they are a by-product of the concurrent growth in the number of electronic records series newly described in NARA's online Archival Research Catalog (ARC), most of which are series not accessible through AAD. For example, a large proportion of NARA's digital series of survey data and statistical data are now described in ARC records. They are not included in AAD because survey data are not suited to record-level access and most of the historic statistical data are not in sufficiently high demand at the individual record level. Nonetheless both are major categories of NARA's accessioned electronic records, so requests about them and orders for copies of files of this data type are a constant. Responding to inquiries about these records requires that reference staff be familiar with or learn about the characteristics of this type of digital data records and know how to utilize the agency-prepared technical information for the records, as well as the electronic records program's in-house administrative databases, as sources for the information necessary for responding to requests.

One of the more intriguing developments since we began offering no-cost online accessibility to a selection of electronic records is that approximately one-third of the fee orders for copies of full files of electronic records on removable media that the electronic records program has processed since AAD have been for copies of files that are freely accessible in AAD. AAD was designed to facilitate online record-level retrieval and supports limited downloading functionality. Except in the case of files having fewer than 1,000 records, it does not support the downloading of full files from a single query. Many researchers have told us that they first "discovered" that NARA preserved a particular series of electronic records when they "found" them in AAD. They contact us because they wish to order a copy of the file(s) for use with their own hardware and software for research or commercial purposes. In other words, one result of offering a new access service for electronic records is that we have come full circle, albeit traversing a much larger circle, back to the form of access to electronic records that NARA has been offering for over 30 years: reproductions of files. This is only the most recent example of a phenomenon we have observed previously in the evolution of NARA's electronic records program. Using new technologies to introduce new user services enhances and expands upon the services offered or communities served, rather than, at least in the short term, replacing traditional services and expectations for them. It is also a reminder that access to records, especially those that are digital, can be accomplished in many different ways, a recognition that the multiplicity in modes of access to identical archival records is one of the affordances of records preserved on digital media.

New Skills

On the most general level, the archival skills required for responding to user expectations are the same, whether records are analog or digital. However, just as responding to requests for information from or about different types of analog records requires that the reference archivist

be conversant with the unique characteristics of specific types of analog records, and have at least a general comprehension of the subject matter of the records, so too it is essential for the archivist charged with providing reference services to have knowledge about and experience related to digital records, a thorough understanding of the variety of modes of access offered for the records, and the technical characteristics related to these modes. That knowledge and the skills for applying it comes from and is rooted in the processing of digital records that must precede their availability for reference or access services. Likewise, the capture of technical information about the records at the file level at the time of processing, the availability of robust descriptions of digital records series, and the preparation of supplementary reference reports or other finding aids, all enable the reference archivist to prepare informed replies.

The knowledge required for all archival reference service, including knowledge required in a digital era, is comprehensively outlined in the Academy of Certified Archivists' Role Delineation Statement and features:

- knowledge of policies and procedures governing access, reference services, and records reproduction;
- knowledge of laws, regulations, and ethical principles governing copyright, freedom of information, privacy, confidentiality, security, and equality of access;
- knowledge of research strategies, needs, and past and current research interests and trends of the communities of researchers to be served;
- knowledge of appropriate reference strategies for the varying types of holdings, formats, media, and user needs;
- knowledge of the subject areas of the holdings, and ideally, how they relate to holdings in other repositories; and,
- knowledge of techniques for expediting the handling of repeated requests for the same or similar topics including using reference files, reports, and frequently-asked-questions (FAQs) materials for online or traditional distribution.[71]

To these generic knowledge areas, we have identified throughout our scenario the manner in which archivists working with digital records rely upon technology as a tool for all of the work that they do. Once they understand how to use the technology available to them, they also need to have a working knowledge of the technical characteristics of the kinds of digital records they are working with or for which they are providing reference services, and the implications of the particular characteristics. Learning to use the appropriate technologies for the archival functions to be performed is an essential skill for archivists in the digital era.

[71] Paraphrased from "The 2003 Role Delineation Statement Revision," found on May 8, 2006 at the homepage of the Academy of Certified Archivists, http://www.certifiedarchivisits.org/html/RoleDelineation.html.

Becoming as "comfortable working with electrons as . . . with paper,"[72] projects a focus on media rather than the recorded information. Setting aside the notion of comfort levels with varying media allows embracing the reality that the universe of valuable primary source material is analog, digital, or a combination of both. To process, preserve, and provide access to digital archival materials requires using the tools of digital technologies. The lines between technologists and archivists are not blurring; archivists, as most twenty-first century professionals, use technology to do their work, while technologists develop the tools with which to work. Perhaps more to the point of a real challenge: archivists need to understand that there are a multiplicity of digital technologies, and thus a multiplicity in the types of digital records, and undoubtedly, a multiplicity of tools for managing them. Unlike the analog world where the authenticity of recorded information and the medium on which it is recorded are one and the same in perpetuity, maintaining the authenticity of digital records is not bound to specific media over time. The media on which records are created and originally used ordinarily will not be the same type of media on which they are preserved nor even necessarily the type of media on which they are provided to researchers. Further, some types of digital records are extensions of analog record types, while other types of digital records are wholly unique to the digital environment. Even when digital records are an extension of previously analog records, computing technology usually supports a multiplicity of ways to manage the records. The most efficient or appropriate application in any given circumstance may not mimic the "way things have always been done," nor does it need to. The archival enterprise has become more complex and multi-faceted; so too have all of society's institutions.

Throughout our scenario, "evolution" has been a constant theme that has no end point. The National Archives could not, in the twenty-first century, be offering online access to even the small subset of historically valuable electronic records now in AAD if NARA hadn't tried to keep pace with technological evolution, verified and gained intellectual control of the records when they were transferred into archival custody, and preserved and routinely migrated them to new media. Because of all of that, these historic digital data can be imported into contemporary desktop applications, enhancing the value of their archival preservation. NARA staff learned about old and new media by working with them, all the while preserving the immutability of the digital records themselves.

In the scenario we commented only briefly on some of the newer types of electronic records that NARA is now accessioning and for which it is offering reference services: digital photographs and narrative textual digital records. They serve as examples of how NARA is adapting what has been learned over the years managing analog still photo, textual records, and digital data records to additional types of digital records. Some of the skill building to accomplish this work has been so incremental that it is hardly perceptible, except in hindsight.

---

[72] Richard Pearce-Moses, "The Winds of Change Blown to Bits," presentation at the Society of American Archivists' 69th Annual Meeting, New Orleans, August 19, 2005, found online at www.archivists.org/presidential, May 1, 2006.

Our scenario describes the interdependence of the processing of digital records and reference services, suggesting what may be, in some cases, a new form of collaboration among the staffs responsible for various functional activities. Because digital records are not tangible objects, informed reference service is totally dependent upon the intellectual control and other products of archival processing. It cannot be offered, whether online, near-line, or offline, until processing has occurred. While archival processing has always been a desirable pre-step to offering reference services for analog records, and as a practical matter a necessary first step for providing access to large and/or heterogeneous collections of analog records, it is essential for digital records.

The scope and scale of the custodial electronic records program at the National Archives, including its AAD resource, reflect the national nature of NARA's mission and responsibilities. Yet the standard processes NARA has adopted for accessioning and preserving digital records, and NARA's development of optional access modes for digital records, could be adopted by archival institutions of varying scopes and scales. NARA's current tools for managing electronic records are, after all, the building blocks for the expanded capabilities expected of the ERA. The NARA solutions to managing electronic records, past, present, and future, offer examples of "how to" but are not the only examples that might be modeled. Much of the work at the academic data libraries and archives, which exist at many research universities and colleges throughout the U.S., has been innovative and path breaking. These institutions are home to local experts in digital preservation and access to digital materials. They may be potential partners for some state or local government archives or campus archives seeking to enter the digital materials arena. More information can be gleaned from the website of their professional association, the International Association for Social Science Information Service and Technology (IASSIST) at http:www.iassistdata.org.

Finally, as we contemplate defining the skills archivists need in this newest period of the digital era, we would do well to return to some of the earlier work on this topic. For example, almost fifteen years ago, Linda Henry shared with the profession her perspective on the potential challenges for analog archivists who might be considering becoming digital archivists in "An Archival Retread in Electronic Records: Acquiring Computer Literacy."[73] The passage of time has not lessened her essential message.

[73] Linda J. Henry, "An Archival Retread in Electronic Records: Acquiring Computer Literacy," *American Archivist* 56 (Summer 1993): 514-521.

10. *One County's Attempt to Move from Zero to Digital in Record Time*

Rich Dymalski and Jerry Kirkpatrick

On January 3, 2006, the Arizona State Library, Archives and Public Records (ASLAPR)[74] forced most government entities in the State of Arizona to grapple with the issue of responsibly managing e-mail records when they issued the "Guidelines for Managing Public Records Sent and Received Via Electronic Mail." Our response at Maricopa County was to form a work group to research the situation more fully and propose a solution to our County leadership.

What we learned through the process is that there is a fundamental shift underway in the way public records are managed. What has primarily been a paper-based activity is now quickly becoming an electronic-heavy endeavor. What we needed was a synergetic partnership between Records Management and our IT departments. What we found was that we had to forge this partnership almost from scratch, and with great reluctance from some in both departments. We would like to share with you what we have learned from both our research and our undertaking in the hope that our journey will help others facing similar challenges. While the digital age may have dawned recently, the practice of records management remains largely, I believe, in the pre-plastic paper era.

Scenario: Prelude

The task of government records management changed dramatically in the State of Arizona on January 03, 2006. That is when the Arizona State Library, Archives and Public Records (ASLAPR) issued the new *Guidelines for Managing Public Records Sent and Received Via Electronic Mail* for all state and local government agencies and political subdivisions in Arizona. These guidelines require three main things for e-mail Records Management: capture the body of the message, metadata, links and attachments; implement a recordkeeping system that is both secure and accessible; and retain each e-mail record for the required retention time period specified for a record of that type (record series), from an approved records retention schedule. These are laudable goals, but the difficulty lies in the fact that these guidelines amount to an unfunded mandate.

Where We Were Before the Guidelines

I swear to you right up front that Maricopa County is not a sleepy, backwater type of county. We have a population approaching 4 million people. Our county is the fourth most populous county in the nation (larger than 21 states and the District of Columbia.) Our County mission is to provide regional leadership, fiscal responsibility, and the necessary public services so that residents can enjoy living in a healthy and safe community.

To be honest, the ASLAPR E-mail Records Guidelines showed us that our county was unprepared for the challenges, financial commitment, politics, creativity, and just plain hard work required to bring Maricopa County from Zero to Digital in Record Time. In 1996 the county

[74] Editors' note. The views presented in this case study and the follow-up report included below are not necessarily those of the Arizona State Library, Archives and Public Records, a sponsor of the colloquium.

had a population of approximately 2 million citizens with a government workforce of about 13,000 employees in 40 business units (departments). Paper ruled, and it did so absolutely. Paper was the normal mode of business, and it was fairly manageable from a records management standpoint. From an IT perspective, we did have plenty of technology: major and centralized data centers (five), many technology operating environments (ten), many network environments (five), several Office-like packages (two), and several e-mail systems (three). IT at the county level was largely traditional, transaction-based application systems, with technologists performing normal backup for recovery. E-mail was a novelty; as a means of communication it was costly, limited, and informal.

Fast forward ten years to 2006, and Maricopa County has a population very fast approaching four million citizens, while our workforce has grown to 15,000 employees in 60+ business units (departments). The county has one records manager, and each department has its own records coordinator. IT has seen the greatest change, with 20+ IT departments, data centers, and widely dispersed PC technology, a single integrated network infrastructure, technology operating environments (four moving to two), MS Office and MS Exchange as standards. E-mail is now the life blood of county communications and the web has replaced many paper-based business operational processes. Our business and citizens are now served by web, e-mail and content-based solutions, technologists still perform normal backup for recovery, and Electronic Document Imaging Systems (EDMS) are poised to rapidly replace the remaining hard copy paper. Paper no longer rules absolutely, and information exchanges with external entities are being reengineered into electronic form.

Where We Are Now

One of the first things we did as a county was to read the guidelines. Once the County Records Retention Officer did so, he contacted the Office of the Chief Information Officer (OCIO) for assistance. After reading the ASLAPR E-mail Records Guidelines, we all gasped a collective, "Oh my Government!" and shook our heads in disbelief. We knew we needed to get down to work. The team has formed the nucleus of what would become the Maricopa County E-mail Records Project Work Group (MCERPWG). What? (Does every acronym have to be catchy?) What we were forced to do was forge a strong, flexible partnership between records management and IT. These two entities and practices have not always seen the need to play together nicely in the same sandbox. But, records management in the digital age has become information management (content management). The two endeavors need to find a synergetic fusion. Both the Director of Materials Management and the CIO took up the challenges of lending their departments' staff, time, and resources to address the needs of a shared response.

The first thing we had to learn was how to speak the same language, or at least one that shared some of the same definitions for the same words. We learned to establish the differences between delete versus purge, archive vs. store, and so on. I believe the vocabulary learning curve for the OCIO / IT folks was easier than it was for the County Records Manager. One of the biggest challenges facing traditional (nondigital) LARMs will be the need to become familiar with,

then comfortable with, then fluent in, IT-ese. Without language proficiency, there can be no knowledge exchange—only assumptions.

The second thing we did was begin to get the word out to only those in the county that would be most directly impacted. The last thing the team wanted was to cause chaos and panic within the County's 15,000 employees. We met with the department records managers and PC LAN managers for a joint information session. This allowed us to start the melding of two different approaches to the same commodity. We will continue to build upon this teamwork approach as the county moves ahead with the implementation of necessary changes, and beyond.

Next, we began to draw together a team of county experts in the areas needed by the work group, and for a resource base. We enlisted the four County e-mail Administrators. (Our county maintains four "separate but equal" e-mail systems.) These four worked together to start reviewing numerous vendor presentations, and sorting through the technical requirements of competing e-mail system solutions. The County Records Manager had limited experience with IT and technology oriented solutions prior to this. Now, he is beginning to acquire a taste for it—as will all LARMs in this digital age. We then sequestered the services of our Maricopa County Attorney's Office, Civil Division, to help us through the legalities of the guidelines themselves, the complexities of public records law, and the debate between the need to properly archive e-mails as public records, yet retain deleted e-mails for internal investigative purposes. There is disagreement between our legal minds as to what to keep, and for how long (value versus discovery). One of the last crucial pieces was requisitioning the assistance of our Office of Management and Budget. Being a government entity, nothing is possible without the funding and buy-in of the budget office. The two OMB members have been essential in preparing the cost analysis that gives our report its financial weight. As Chairman Mao once said, "A revolution is not a dinner party." Nor, is an undertaking of this importance and scope a "free lunch."

As our team began to come together, the assistant director of the OCIO drew up an excellent game plan. The strategy for the E-mail Records Project Work Group grew into a multistep approach: conduct research, discuss alternative solutions—technical and business, determine impacts and liabilities, meet with the County Attorney and counsel for legal opinion, analyze the costs of solutions and impacts to business operations, technology service, and support; and then present our findings and recommendations to County leadership. We had an excellent roadmap, a crack team assembled, and it was only February 01, 2006. A fair amount of work for one month's time, but the real work was just starting.

As part of the strategic plan, we also defined the following areas of concentration: available e-mail archiving and technical support solutions, legal implications and inconsistencies; what other Arizona government entities were doing; similar legislation and solutions in other states; employee knowledge, practices, and training; current county e-mail and records management policies. We conducted our research in a variety of ways, including meetings of the full work group, small subgroup meetings, brainstorming sessions involving one-on-one calls, e-mails, website searches, and the always reliable "all-nighters."

At this point, we are feeling good about our research work. The report and proposed recommendations are almost finished, and the hard work is paying off beyond our wildest expectations. The research of this work group has been able to pinpoint areas of inconsistencies in the practices of our County's records and information management, and target specific solutions to these areas of need. We have arranged for several forums where we can share our research and findings with other government entities in Arizona, and are looking for ways to reach those grappling with similar complex problems beyond our state. It is our hope that the work done, the research conducted, the findings put forth, and the solutions proposed will lead to better government in our county, more efficient and effective policies and practices, and the better service of our citizens.

Where We Are Going

"We need to crawl before we can walk, walk before we can run." This has become the motto of our research findings and proposals. We have so much ground to cover as a county that it would be both discouraging and impossible to put forward anything but a measured, phased solution. Here is a summary of our findings and our recommendations:

- Our employees and business operations cannot take a large leap in how they have to work—we need to start with the basics (type, then category, then classification, then retention period).
- The culture of government is based on a manual, paper dependent operation—we must change the thinking, the culture, the disciplines and the processes to fit the electronic world we are creating.
- We do not have the authority to change—we must get executive management approval and backing.
- We are inconsistent in our methods—we must create standards, train those standards, and fund the changes.
- Our employees do not know what to do—we need to introduce the concepts at initial hire in our new employee orientation, train existing staff through our intranet, and require annual re-certification.
- Our employees may not all comply with the changes—we need more than new skills training. We must change the attitude and philosophy of our staff; evolve the mindset.
- We cannot afford the impact of these E-mail Records Guidelines—we must try to change our direction and the ASLAPR mandate from an all-or-nothing approach to a phased solution.

What We Will Need to Get There

The most important thing that we will need to make effective, consistent, long-lasting change on a county-wide level is executive level and top management buy-in, funding, and support for the following:

- change in **competencies** (knowledge, skills, abilities) for hiring through modified job requirements and practices
- change in **culture** (emotion, work habits, psychology, thinking)

- change in **discipline** (laws, policies, standards, procedures, other controls)
- change in **organization** (structure, mission, jobs, responsibilities)
- change in **processes** (activities, functions, tasks, flow, stats)
- change in **technology development** (build into initial system and not try to add on later)

New Skills

Skills we will need from a *LARM* perspective:

- Bring together a team of county experts from various disciplines.
- Pinpoint areas of inconsistencies in the practices of records management and information management.
- Forge a strong, flexible partnership between records management and IT.
- Work past the issues of territoriality and superiority inherent in RM vs. IT thinking.
- Learn how to speak the same language, one that shares the same definitions for the same words.
- Become familiar with, comfortable with, then fluent in IT-ese.
- Become familiar with IT solutions architecture in order to learn how to evaluate IT solutions.
- Start the fusion from two different ways of thinking and approaches to a single commodity driven culture.
- Begin to think "electronically" in the business of LARMs:
    - Why print and file e-mails (electronic mail)?
    - How can a database be tracked on a records retention schedule?
    - How do e-data and paper documents differ in usability?
- Accept ownership of the data, and the responsibility to determine its importance and life.

Skills we will need from an IT perspective:

- E-mail content can be a business record
- Records management includes any type of information communications appliance (file, database, e-mail, IM, video conferencing)
- Records management terms under a paper system do not necessarily mean the same in the electronic world (archive is not backup, disposal is not delete)
- Large number of record series and retention periods
- Allowing only paper or microfiche / microfilm for electronic images, and for archiving of e-mails
- Technology that can handle e-mail records management
    - no solution that can handle the complexity of government rules on record types and retention periods found in our current retention schedules
    - no solution that really incorporates all the required steps in good records management
    - no solution that seamlessly couples itself to other business enterprise applications
- Technologists serve the needs of the business for information management (stewards).

REPORT by E-MAIL RECORDS PROJECT WORK GROUP
Maricopa County, Arizona

OBJECTIVE

The objective of the work group assembled to study this subject is to measure the impact of the Arizona State Library, Archives and Public Records (ASLAPR) E-mail Records Guidelines on Maricopa County and deliver a report to County leadership by August 2006.

INTRODUCTION

The Arizona State Library, Archives and Public Records (ASLAPR) issued a set of guidelines across the State of Arizona for managing e-mails as public records in a web publication on January 3, 2006. The publication caught Arizona records managers, including Maricopa County's Records Manager, off guard. While electronic records and their management as public records have been line items on retention schedules for many years, practice has lagged behind policies. The guidelines came suddenly and assumed compliance as of the date of issuance.

In discussions among County departments following issuance of the Guidelines, many ideas surfaced. Some of them relate to ASLAPR's authority over our departments. Another has to do with the weight of the ASLAPR guidelines. Others relate to how reasonable it is to expect immediate compliance with the issues associated with e-mail and electronic document management: with types and costs of technical solutions, the difficulties of defining business and functional requirements, with training, and with the hope that by doing nothing the issues will simply go away.

The purpose of this report is to make County leadership aware of the issues, to outline the investigative strategies employed to date, and to propose a course of action.

STRATEGY

I. Explore Implication of Guidelines

Jerry Kirkpatrick, the County's Records Manager located in Materials Management, was charged by department director Wes Baysinger to study the issues resulting from the new ASLAPR guidelines. Jerry created a work group representative of the challenges and the County. Membership consists of records managers, IT experts, and others representing staff from every major constellation of our organizational chart. See Exhibit A for a list of the members of the work group.

The largest implication of the ASLAPR guidelines is the fundamental shift from viewing records management as a largely paper-based endeavor to one of an electronic activity. Previously, records management revolved around paper. Each department had one records

manager charged with understanding and adhering to the ASLAPR records retention schedules, and managing a department's paper records. The guidelines now require every County employee to be a records manager, to have a thorough working knowledge of all retention schedules, and to make correct (i.e., "liable") records management decisions for every e-mail that is sent or received.

In Maricopa County, that means that rather than requiring only 60 to 75 departmental records managers to manage departments' paper records according to their retention schedules, ASLAPR now requires each of the our 15,000+ employees who send or receive e-mail to manage those records based on the County's records retention schedules. This is definitely a new risk the County will need to explore.

Members of the work group have had questions about ASLAPR—who they are and how their authority can affect Maricopa County. The ASLAPR's Board is composed of four members of the Arizona Legislature. Its director is responsible for the preservation and management of records. The main purpose of the ASLAPR is to establish standards, procedures, and techniques for effective management of records (ARS §41-1345). An ASLAPR approved records retention schedule legally provides for the proper destruction of records. ASLAPR has no enforcement division, but the ARS (irrespective of the recently issued Guidelines) require that public officials adhere to ASLAPR records management requirements (ARS §41-1346, 1347).

ASLAPR published the *Guidelines for Managing Public Records Sent and Received Via Electronic Mail* on its website on January 3, 2006. The key definition to consider is the one for "record" as stated in ARS §41-1350.[75] The guidelines cover a number of areas. Highlights include requirements for managing e-mails as public records, policies to cover the same, and employee training. See Exhibits B and C for detail.

The County's Electronic Mail Policy, A1608, dated February 1997 and revised August 2001, meets almost 95% of the ASLAPR Guideline requirements. This policy and our current retention schedules reflect the main points of the guidelines. The problem is that County employees have not been trained on the content nor have they been provided a practice or method to comply with these policies. Therefore, compliance on a County-wide level is poor, at best.

II. Research

Throughout the course of the work group's research, it has become clear that e-mail use in conducting business in both the private and public sectors has exploded. There is no going back.

---

[75] The definition is nearly 150 words long, including almost 50 words detailing what is not a record.

[2] Storing e-mails to either .PST files or onto CD or DVDs was considered but excluded, in part because of difficulties relating to retrieval, retention, and archiving.

As a result, businesses providing hardware and/or software solutions for e-mail storage are among the fastest growing in information technology today.

The work group struggled with interpreting the guidelines given the impact on the way the County currently conducts business. We looked to the MCAO Civil Division for assistance. The MCAO Civil Division attorneys on the work group drafted a memo addressing the legalities of the ASLAPR guidelines and the extent to which they may apply to Maricopa County. At the time of this report, no ruling on the memo has been issued.

What the MCAO Civil Division has made clear is the inclusion of e-mails as public records. The role of e-mails in legal and judicial proceedings is accepted practice. The fact that Arizona has one of the broadest definitions and practices of a "record" is widely known. The statutes require that the County strictly adhere to all ASLAPR approved records retention schedules (ARS §41-1346.8). Both the ASLAPR and Maricopa County retention schedules state that e-mails can be records and must be properly retained, if such, according to approved retention schedules.

According to ASLAPR, the guidelines apply to almost all public agencies in the State, although we are waiting for the MCAO Civil Division opinion on the true scope. We have contacted other city and county records managers, and have found that all are grappling with the guidelines, but in differing degrees. Maricopa County was part of a panel discussion on e-mail management at the Arizona Digital Government Summit held in Phoenix in May, 2006. The guidelines and the complexities of compliance were briefly, yet vigorously discussed.

Discussions on e-mail management, retrieval, and storage are happening globally. Businesses across the U.S. must confront the issues and find solutions thanks to the Sarbanes-Oxley Act of 2002 (SOX). Although SOX does not apply to government, governmental bodies are affected by the requirements to manage electronic records. We have found the following states to be leaders in the field: Delaware, Florida, Kentucky, Maine, New Jersey, New York, Ohio, and Texas.

Preliminary research indicates that solutions to e-mail records storage generally fall into four categories:

- Print all e-mail records onto paper and file with other paper records;
- Store e-mails within the current mail system (i.e., Outlook), or the new version, Exchange 2007;
- Store e-mails in a third-party archival or e-mail management system;
- Manage and store all records and documents (not just e-mails) within an integrated electronic system known as an Electronic Document Management System (EDMS).

The ASLAPR guidelines require that employees be trained in e-mail records management. Online training broken down into four thirty-minute segments is the best way to train a large number of employees consistently, efficiently, and in the shortest possible time. More in-depth training would be required of all department records managers.

III. The Ostrich Approach is not an Option

After a series of discussions, the work group concluded that while departments are not interested in assuming unnecessary risks, especially associated with costly technology solutions, to do nothing while waiting for others to address the implications of the guidelines puts the County at risk.

This report does not address legal risks of failing to comply with ASLAPR directives. Any discussion of risks associated with partial or non-compliance to the ASLAPR guidelines can be addressed by the the Maricopa County Attorney's Office, Civil Division. However, members of the work group are aware that the statutes mandate the following penalties for records management non-compliance: class 2 misdemeanor for not promoting a records management program (ARS §41-1346.C); class 2 misdemeanor for imaging / scanning records without prior approval (ARS §41-1348.D); class 4 felony for an officer destroying records prematurely (ARS §38-421.A); and class 6 felony for a non-officer destroying records prematurely (ARS §38-421.B).

The County's records management program needs to be defensible from a legal standpoint. Credibility is essential. There are three things that would add credibility to our County's records management program:

- the existence of policies and procedures that fully address and define the issues and practices of sound records management;
- employees that have been properly educated and trained on the essentials of records management and the particulars of e-mail records management;
- consistency across the County in the practice of our records management program.

These three areas can be affected positively if we follow the recommendations and proposals herein.

There is always a risk that attention of the public and/or the media could be negative if they perceive that the County is neither supporting nor practicing acceptable access to all public records. (Records management is essential to public records access.) We all know the importance placed on the public's right to access its records. While ASLAPR has no enforcement division, it has and will step in rigorously and tenaciously at the request of the public.

Costs and Formula

The County's mission and practice of fiscal responsibility are well-known. While our research has included solutions with large price tags, our response should be conservative and phased to spread out costs.

The following formula will be incorporated from this point on in any costing of solutions.

- The work group has engaged in a consistent methodology in developing the numbers for the cost benefit analysis and report. We have based the figures on percentages of the overall County figures, including extrapolating figures attributed to the four e-mail administrators.

- We have determined that a conservative average of 30% of all e-mails received will qualify as "records." (This figure has come from NARA and ARMA, major records management organizations.)
- Five years is the average retention period for our records, based upon our current, approved records retention schedules.
- In the course of our investigation and research, there has been a request that the County keep "deleted" e-mails for one year, up from the current 28 days. (This will provide a better tool for investigations and legal evidence.)
- We have accounted for a 25% growth annually in County-wide e-mail usage.

IV. Weigh Alternatives

Stand Still

Both the ASLAPR guidelines and County Policy A1608 provide for a compliance solution of printing and filing all e-mail records. What we have discovered, however, is that this solution is both counter-intuitive and the most expensive option. The county processes 3.8 million e-mail messages monthly for 16,000 mailboxes. (This figure does not include the MCAO.) The average e-mail message is 90 KB. If printed, this 90 KB will take two to three pages of paper. Printing only record e-mails will consume 2,280,000 pages per month. The estimated total combined cost for paper, laser jets, toner, and maintenance will be **$186,276** per month, or **$2.24 million** annually. The cost of manually boxing the paperwork and off-site storage will run **$207,349** annually. (The cost of manually filing the paperwork, filing cabinets, and floor space would increase this figure, and has not been calculated.) The ten-year total cost for a paper solution to e-mail will be **$28 million**. We are not recommending this approach, although it is an option.

A dollar figure of some size to consider with this option is the cost of lost lawsuits due to the fact that the County would encounter considerable difficulty in managing an additional 2.2 million pages of paper monthly.

Walk

Every e-mail system solution we researched involved the purchase of more electronic storage / servers. Without the purchase of additional e-mail storage space, the only way that Maricopa County can comply with both our current approved retention schedules and the ASLAPR guidelines is by printing and filing e-mail records. (See "Stand Still," above) The County cannot make even the smallest progress toward e-mail records management without increasing the size of the mailboxes for its employees. Without further research and investigation, and a determination from the County leadership on what their requirements and needs are, we are not able to provide an accurate price for interim storage needs.

There are, however, seven steps that we can take immediately, which will incur minimal cost to the County. (These seven steps are listed in the **Recommendations** section under **Step One**, below.) Much of the work will be done by the County Records Manager and the E-mail Records Project Work Group, and will provide the County Records Management program much

needed credibility, consistency and compliance, while providing legal leverage for the County in the process.

Run

With this approach, the county would begin training our employees on paper and electronic records management. This training is essential for the success and progress of the County's records management program. The recommended approach to training is four 30-minute online modules that all County employees would be required to take and pass for certification. All department records managers would be required to take more in-depth, classroom training. However, no employee training can begin until we have enough mailbox storage for the additional e-mail records retained.

Fly (two options)

Option One

Along with employee training, the County would upgrade to Microsoft Exchange 2007, the next version of Outlook, scheduled for an early 2007 release. Exchange 2007 will require additional storage / servers. (Whichever system we choose, it will require additional storage.) But, Microsoft Exchange 2007 can meet almost all of the requirements set by ASLAPR "out of the box." And, the County already owns it. Exchange 2007 includes a backend advanced find tool, much like advanced find in Outlook; and it covers all of the Exchange system.

There will be an additional charge from Microsoft to utilize the e-mail management capabilities, similar to additional licensing. This additional charge will cost the County **$2.6 million** over the next 10 years. Without these optional e-mail management capabilities, *Exchange 2007* will be an ineffective e-mail management solution.

Option Two

Along with employee training, the County could decide to pursue an Electronic Document Management System (EDMS), enterprise-wide. The Maricopa County Clerk of the Court currently uses the OnBase system, an EDMS, and the finance department is currently implementing OnBase. Several County departments are considering implementing the OnBase system, including the Assessor's Office, the MCSO, Air Quality, and others, and we have a new contract for OnBase.

The real strength of an EDMS is that it allows all types of records to be managed, not just e-mail. With an EDMS, the potential exists for a "paperless" government. The enterprise-wide price tag for OnBase is approximately **$3.5 million** for licensing, which does not include removing costs already paid by the Clerk of the Court and other other county departments for OnBase licensing. OnBase is a software solution, and will require additional hardware storage to support it.

V. Recommendations

Phased Approach

After all of our research, we believe the best approach would be a reasonably conservative approach, phased in over time. Here are the recommendations:

Step One:

1) Revise and gain approval for revisions to the A1608 E-mail Policy and the A2100-series Records Management Policies. Revision would be based on providing a proper program, consistently applied across the County. The revisions would include any aspects of the guidelines that the MCAO Civil Division would not question, and aspects that are already a part of our policies and/or our retention schedules. The revisions would be specifically tailored to match the guidelines. (There are currently seven A2100-series policies that have been written and have being reviewed by the work group and the department records managers and PC LAN managers.)
2) Develop a working definition of a record that would be easily understood and practiced by our employees. (This has been developed, and is contained in the A2100-series policy revisions.)
3) Rework the *Maricopa County, All Departments, Administrative Records – General Schedule* with a focus on simplifying and condensing the number of records series, line item numbers, and retention time periods. Our focus would be on arriving at five main categories: short term (6 months maximum), one year, three years, five years, and seven years.
4) Meet with ASLAPR, only after the retention schedules have been reworked, to discuss our research findings and our need to streamline our retention schedules.
5) Maintain current retention schedules, policies, and practices, and lobby the Arizona Legislature to overturn the ASLAPR E-mail Guidelines. We could enlist the County's government relations department to work with the legislative process to overcome this issue.
6) With the help of the County's government relations department, begin to meet with other State agencies, government entities and organizations to discover how we can work together to make the ASLAPR guidelines more reasonable, practical, and user-friendly. We should begin by meeting with the ASLAPR Director and division directors to determine their level of commitment to the e-mail guidelines, and any possible flexibility. We could then follow up by meeting with the four ASLAPR Board members, all members of the Arizona Legislature: President of the Senate; Speaker of the House; one additional senator; and one additional representative.
7) Start a thorough search of the County's e-mail storage needs. Then, purchase the recommended amount of e-mail storage / servers to provide our employees with much needed mailbox space, provide the County with some much needed time to consider its options, and allow us to begin training employees with the expectation that employees will have the necessary tools to practice sound records management.

Step Two:

1) Begin the recommended online training to provide records management / e-mail management training to all Maricopa County employees. We recommend targeting the department records managers and the department PC LAN managers for more in-depth classroom training. The curriculum could be developed by a sub-group from this work group based upon a current e-mail training out of MCDOT, and should be based on the A1608 E-mail Policy, and the A2100-series Records Management Policies. The cost for developing the training, and time lost while all County employees take the training will be **$831,600** for the first year, with a 10-year total of **$1.3 million**.
2) Add a segment on records management to the County's new employee orientation so that all new employees are properly equipped to comply with our policies and practices.
3) Have the MCAO Civil Division consider the legality of including voice-mail, databases, instant messaging, VOIP, teleconferencing and video-conferencing, and other new communication technologies on the fringe, under the ARS definition of record. ASLAPR plans to include these communications tools under the definition, and we would do well to figure out a legal stance early.

Step Three: Two options

Option One: Make full use of the Microsoft Exchange 2007 system to serve as both the County's mail system, and provide a major tool for management of e-mail records. The County will need to purchase additional storage / servers to ensure the success of the Exchange 2007 system. The 10-year cost for licensing the e-mail management tools, hardware, storage, support services, and additional staffing will be **$23.9 million**.

If we decide against the recommendation to archive all deleted non-record e-mails for one year, and maintain the current practice of archiving deleted e-mails for 28 days, the ten-year cost for licensing the e-mail management tools, hardware, storage, support services, and additional staffing will be **$17.9 million.**

Option Two: Maricopa County could adopt an Electronic Document Management System (EDMS). The OnBase system, currently being pursued by several of our departments, is the best possible solution to both e-mail management, records management, and electronic content management. There are already some legal precedents for including databases, instant messaging, voice mail, video conferencing, and so on as records, and the OnBase EDMS seems to have the potential to manage these other communications tools. An EDMS will allow for a more paperless future, and provide for better records and document management.

We currently have a contract that addresses implementing the OnBase system as the County-wide records management solution. The hardware, storage, support services, and additional staffing costs will be **$2.3 million** more (over a 10-year period) than the Exchange 2007 costs for e-mail management, but we have not deducted costs already paid by other county departments

for OnBase licensing. The main benefit of the OnBase system would be having the e-mail records and all other electronic records for the County in one central repository.

VII. Future Areas of Attention

During the course of research conducted by the work group, we have discussed several areas that will warrant further attention and action by the County. These are, very briefly: the use of personal vs. department drives for storing records; the need to institute a data classification policy and practice; scheduling / tracking county databases; inventorying all County records in electronic format; and the County's need for longer retention of electronic records for forecasting and research vs. listed retention times on records retention schedules.

FINDINGS

There is no question that Maricopa County is an award-winning county capable of sound, conservative, yet creative solutions to the large problems facing us in this digital age of government and business. The ASLAPR guidelines have grown out of the need to address new electronic realities. Records management that was once paper-based is now information management that is electronic in nature. This shift will put some strain on our records management program and practices which are inconsistent in an age when much attention is focused on the public's right to access its records, while also protecting one's personal information. Consistency leads to increased credibility, which leads to increased leverage, respect, and positive attention. Consistency is possible. Our current records management department is the place to keep the management of both paper and electronic records. Maricopa County has policies in place that need revision and updating to meet our needs. Our records retention schedules will need to be streamlined and simplified. Training our employees on proper records management makes sense, and will reap large benefits—one of which will be employee satisfaction at having the necessary tools, training, and skills to perform the requirements of the job. Electronic tools like increased storage, and either Exchange 2007, or the OnBase EDMS will allow us to meet the new challenges and provide increased customer service to both the public and our own departments.

EXHIBIT A

E-MAIL RECORDS PROJECT WORK GROUP MEMBERS

Nothing would have been accomplished without the incredible amount of time, energy, and brain power devoted by Rich Dymalski, Troy Geis, and Sarah Shew. They have been the guiding and driving force of the work group. Also, there would be no cost benefit analysis without the many hours of hard work Daren Frank has contributed to bringing together the figures needed for this report. Finally, the work group has been ably overseen and guided on a department level by Wes Baysinger, Director of Materials Management.

It was important from the beginning that the work group be truly representative of the challenges and the County. The group is a good fusion of records managers and IT experts, representing every major sphere of the County organizational chart. There would be no substance and depth to this report without the knowledge, experience, and commitment that was provided by its members. The members of the work group include Rich Adams, Rachel Alexander, Gina Althof, Chuck Brokschmidt, Martin Camacho, Casey Carpenter, Don Colvin, Mary Cronin, Lynda Cull, Rich Dymalski, David Eisenstein, Cory Farrell, Keely Farrow, Daren Frank, Troy Geis, Andro Gonzalez, Stephen Hamman, Earl Hinkle, Joel Kodicek, Marc Kuffner, Richard Lemon, Cathy Lucero, Mary Lee Madison, Raymond Maiorana, Ward Maeser, Darrell Mills, Mark Murphy, Lisa Nash, Mohit Parnami, Norma Preciado, Cynthia Robinson, Marty Scott, Sarah Shew, Lisa Stelly Wahlin, Amy Thomas, Joe West, Kevin Westover, and Elizabeth Yaquinto.

The following County departments are represented by the E-mail Records Project Work Group: Assessor's Office, Facilities Management, Finance, Human Resources, Human Services, Library District, Materials Management, MCAO, MCAO Civil Division, MIHS, Office of the CIO, Office of Management and Budget, Planning and Development, Public Defender, Public Health, Public Information Officer (Communication Office), Public Works, Recorder's Office / Elections, Sheriff's Office (MCSO), Superior Court, Training and Development, and Transportation (MCDOT).

EXHIBIT B

WHAT THE ASLAPR GUIDELINES REQUIRE

1 - 3. Defines when e-mail is a public record and which components are included

4. **Requires** agencies to establish policies and procedures for managing e-mail
   Compliance: A1608, 1605
5. **Requires** employee training in awareness and management policies and procedures
   Compliance: Poor
6. **Requires** agencies to make e-mail records available under the public records law
   Compliance: Inconsistent throughout County
7. Permits deletion at any time of nonpublic record e-mail messages
8. **Requires** e-mail records to be retained according to specified ASLAPR-approved records retention and disposition schedule
   Compliance: Poor
   Forbids agencies to routinely delete all e-mail after "arbitrary" amount of time
   Compliance: Reason behind new "Your Inbox is Full" warning message
9. **Requires** agencies to suspend e-mail record destruction relevant to reasonably foreseen legal action, audit, and investigation
   Compliance: Inconsistent throughout County
10. **Requires** annual reporting of destruction of public records "without … value" via records destruction form
    Compliance: Paper records – inconsistent; e-mail records – poor
11. Permits several filing options:
    1. Printing and filing e-mail records with other paper records
    2. Use of software to facilitate management and disposition of e-mail records

    Requirements of filing systems:
    1. Secure system
    2. Must capture e-mail content, metadata, links and attachments
    3. Must be accessible as public records
    4. Retention must be specific, based on approved retention schedules

    Forbids use of e-mail system backups for retention of e-mail records
12. **Requires** e-mail records with permanent retention periods to be transferred from e-mail system and stored in either electronic recordkeeping system or another "proper" recordkeeping system
    Compliance: Paper records – inconsistent; e-mail records – poor

EXHIBIT C

ASLAPR E-MAIL RECORDS GUIDELINES

Scope and Responsibility

This document provides guidelines for the management (creation, maintenance, access and use, and disposition) of e-mail messages in accordance with state and federal legal requirements. Public officials and other custodians of public records (hereafter referred to collectively as "agencies") shall preserve and protect public records in accordance with these guidelines and to maintain documentation as evidence that these standards are being met. These guidelines apply to state and local government agencies and political subdivisions in the State of Arizona.

Authority

These guidelines are established by the Director, Arizona State Library, Archives and Public Records pursuant to ARS § 41-1345.A.1. It is promulgated by the Arizona State Library, Archives and Public Records, an agency of the Legislature. This document was initially prepared by a committee acting as advisors to the Director. The committee was composed of records management professionals with representation from State, county, and municipal government. Its purpose was to develop guidelines for managing electronic messages that are public records.

Guidelines

1. E-mail messages created or received by a government employee are public records under ARS 41-1350 if it documents the organization, functions, policies, decisions, procedures, operations or other activities of the political organization.

2. E-mail messages sent by an agency employee in their official capacity using another system (for example, a personal, home e-mail system) are public records.

3. An e-mail record includes metadata (minimally the sender, all recipients, date and time sent, subject), the body of the message, any attachments, documentation of all recipients. If an e-mail record is sent to a distribution list, it must be possible to demonstrate who received the message, not just the name of the distribution list.

4. Agencies shall establish policies and procedures for managing e-mail created or received by the agency, including preserving and filing, access and use, and disposition. Such policies shall address the use of e-mail for sending sensitive, proprietary, or confidential information and shall also address any state or federal legal requirements specific to the agency's work.

5. Agencies shall make employees aware that an e-mail may be a record and shall provide employees training in policies and procedures for properly managing e-mail.

6. Agencies must make all e-mail records available to the public upon request under the Arizona Inspection of Public Records Law (ARS 39-121) during the required retention period, unless the content of the message falls under one of the exceptions contained in the law or in any other statute, regulation, Executive Order, or rule of court.

7. E-mail messages that do not meet the criteria of the Arizona statutory definition of a public record may be deleted at any time, unless they become part of some official record as a result of special circumstances.

8 Agencies must retain e-mail records for the period of time specified on a records retention and disposition schedule approved by the Arizona State Library, Archives and Public Records. Retention or disposition of e-mail messages must be related to the information they contain or the purpose they serve. Agencies may not routinely delete all e-mail after an arbitrary amount of time.

9. Agencies must suspend destruction of e-mail records relevant to any reasonably foreseeable legal action, audit, or government investigation until the conclusion of such action, even if their retention period has passed. Agencies should suspend destruction of potentially relevant records as soon as there is reasonable expectation of such action, regardless of whether a legal notice of such action has been served.

10. Agencies must report the destruction of public records without legal, administrative, historical, or other value to Arizona State Library, Archives and Public Records (ARS § 41-1351) on an annual basis.

11. Agencies have the option of printing and filing e-mail records or may use software to facilitate the management and disposition of e-mail records. Agencies may not use backups of e-mail systems for retention of e-mail records.

12. E-mail records that have permanent retention periods must be transferred from the e-mail system and stored in either an electronic recordkeeping system or another proper recordkeeping system.

Arizona State Library, Archives and Public Records
1919 West Jefferson • Phoenix, AZ 85009
e-mail: rmd@lib.az.us

11. *Mapping Processes in Motion: Practical Lessons from the Experience of Discovering, Visualizing, Analyzing, and Redesigning a Complex Process of Digital Archiving and Dissemination*
Cole Whiteman

In this case study we look at one repository's "process mapping" approach to coping with a large and complex body of technical procedures that change frequently and tend to get out of synchronization with each other. ICPSR has invested effort over the past three years to map its data pipeline process, in order to increase shared understanding among staff about how the process works, and to guide multiple process improvements.

Scenario

The setting for this case study is ICPSR, the Inter-university Consortium for Political and Social Research, which has been hosted since its inception in the early 1960s by the University of Michigan's Institute for Social Research.

ICPSR operates a suite of social science data archives as a service to the research community. What comes into the ICPSR are social science data collections, appearing in a variety of formats and states of assembly; what goes out are standardized, validated, fully documented versions of these collections, posted on the web for today's researchers and archived for researchers of the future.

The Pipeline

How "what comes in" becomes "what goes out" is the product of a data acquisition and processing pipeline that involves many groups, many people, many procedural steps, and many technical tools. The pipeline components have been assembled over ICPSR's 40-plus year history by people dedicated to improvising ways to handle the wide variety of input that the organization receives. Some of the developers of these procedures and tools are still on staff, and they serve as sources of lore about how these components work and why they are the way they are.

The Problem

Over the past decade there arose within ICPSR an appreciation that, while many of the pipeline components have adapted to numerous changes in technology and in user expectations, others have been less responsive; and that this was at least in part because there was no straightforward way for any one person to get a grasp on how the whole pipeline worked, at either a summary or a detailed level.

There had been several attempts over the years to document the pipeline operation, but the results had been less than satisfactory for various reasons: the output documents were incomplete, or lacking in detail, or incoherent, or otherwise unclear. Some documentation attempts were considered "partisan," having been written from one or another group's perspective, without regard for variations in procedures among the groups.

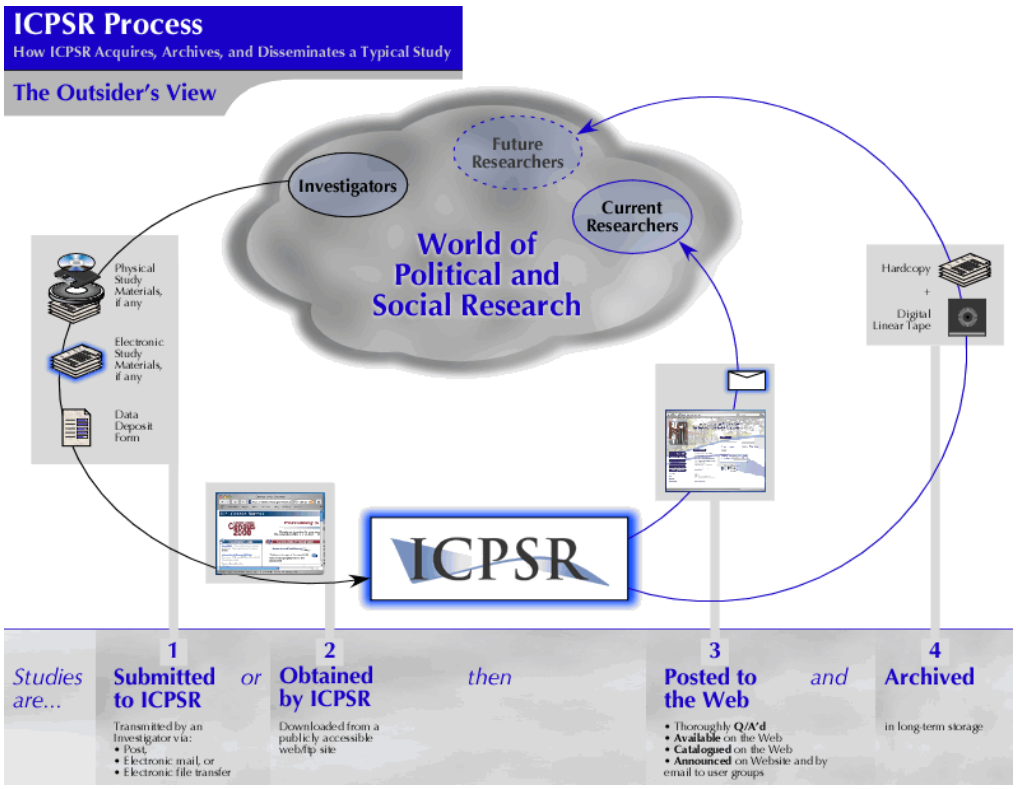
The Solution

Three years ago, ICPSR began to respond by investing in a focused effort to document its pipeline process—systematically, coherently, comprehensively, spanning the entire pipeline from beginning to end, integrating all relevant perspectives, and with results made easily available and comprehensible to all involved staff.

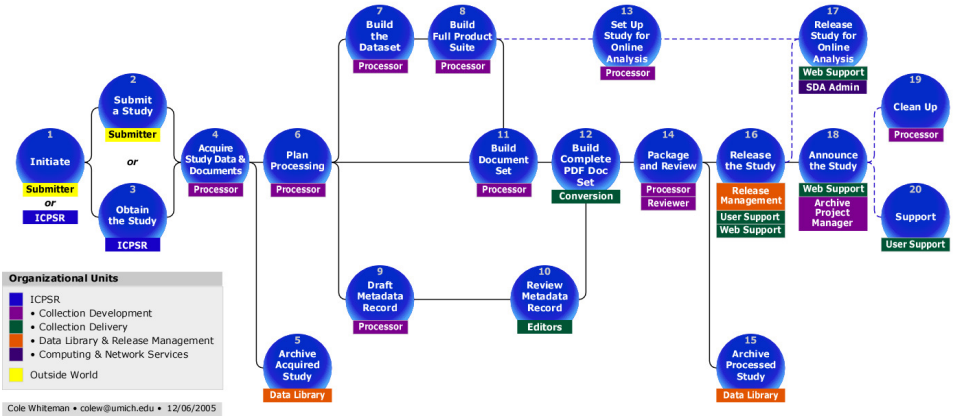
We spent a summer creating an initial set of hierarchical diagrams to depict the pipeline process, with a one-page high-level summary at the top of the hierarchy and a detailed map measuring 3 feet high by 28 feet long at the bottom. We posted these on a corridor wall for staff and visitors to see.

Samples

Here is our diagram of how the process looks to the outside world. The ICPSR operation is an opaque block that receives "raw" studies and delivers web-posted and archived studies.

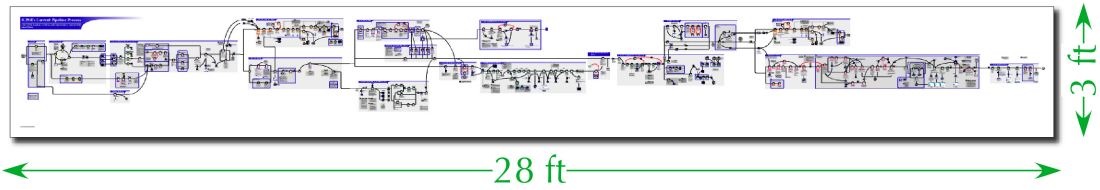


The pipeline behind the scenes, at the top level, looks like this:



Each bubble represents a major step in the pipeline, and has a detailed diagram behind it.

The composite of the detailed diagrams is the one that spans the corridor wall:

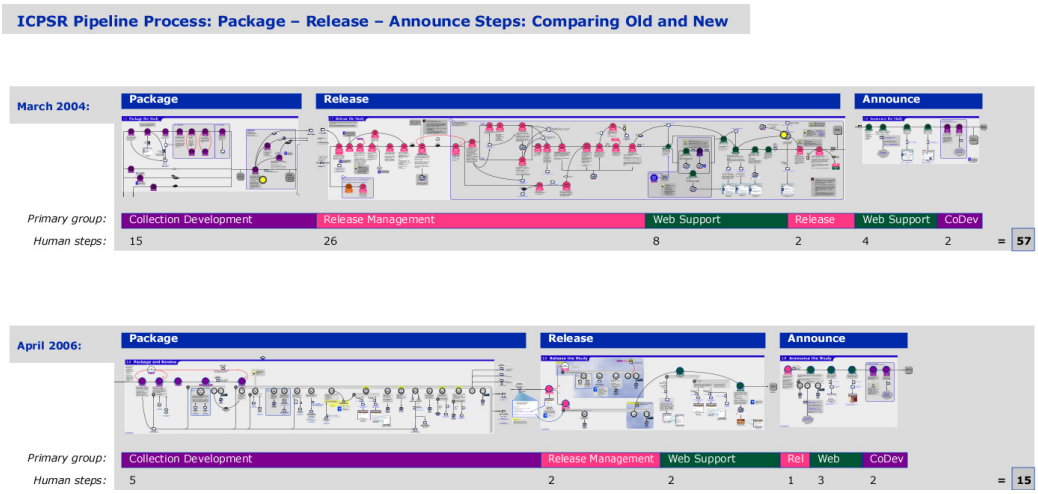


Benefits

This process mapping exercise brought to light a body of distributed and submerged lore about how the process actually works, plus staff desires for how it might work or should work.

The process mapping led directly to a coordinated process improvement effort, with a rotating team devoted to gathering ideas and developing them into proposals. This in turn has yielded a stream of pipeline enhancements, some conservative, some radical. We are now seeing the results accumulate in the form of increased throughput with less effort.

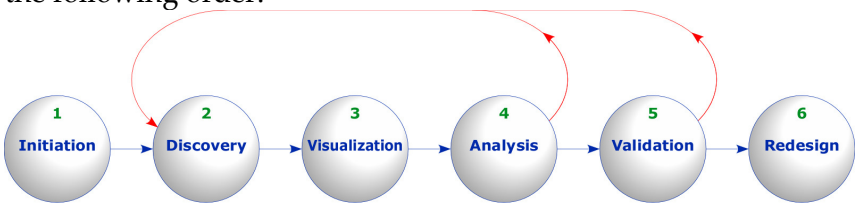
The process map itself has become clearer and more streamlined; we've been periodically updating it to keep up with pipeline automation and other improvements. The map has also been useful as a source for a number of derivative documents we've created for planning and presentation. For example: here's one to illustrate "before and after" a wave of automation on the tail end of the pipeline:



As another spin off from the process mapping exercise, we are now overhauling the internal technical procedure documents on our intranet website, using our high-level process diagram as an organizing principle and as a visual table of contents.

The Mechanics of Process Mapping

How does an organization go about constructing a solid, useful process diagram? We find that this kind of effort involves six activities, which can overlap somewhat but commence in roughly the following order:



1. Initiation

This is where we plan the diagramming effort. We clarify:

- our essential goal;
- our intended audience;
- the scope of the process(es) we're going to describe;
- what tools we'll use to construct the diagrams;
- what format(s) we'll render it in for distribution;
- who will play what role in the effort; and
- any constraints on time or resources.

2. Discovery

This is where we gather information about the process we're describing. It includes talking to people, reading existing documents, and navigating through existing information systems. The "talking to people" can appear informal but is structured around:

- *Interviews*, in which we collect people's stories about how their part of the process works (which we consider authoritative), and how other parts of the process work (which we consider very useful but not authoritative).
- *Reviews*, in which we gather people's reactions to others' stories, and to our visual rendition of these stories, as a way to detect inconsistencies.
- *Discussions*, in which we pull together two or more people who have given conflicting (or apparently conflicting) tales of how something works, so we can resolve the inconsistencies.

3. Visualization

This is where we actually sit down and draw something. We might do early sketches by hand, but pretty quickly we turn to a computer and draft a process diagram using illustration software. Depending on the audience, we might be creating a large diagram to hang on a wall, or a hyperlinked set of small diagrams to post on the web.

For graphics software, we believe that most any of the major illustration packages on the market would serve well. For the ICPSR process mapping effort we chose *Canvas* by ACD Systems because it is cross-platform (Windows and Mac—ICPSR has staff of both persuasions); it integrates bitmap images and vector object graphics, so we can use one piece of software instead of two; it has a library of common graphic symbols, plus a rich tool set for designing new graphics so we can create icons for locally familiar process objects; it can handle large illustration sizes (for hanging posters on walls); and it exports/imports well with PowerPoint and other graphic applications.

As an alternative approach, there are software packages on the market that are specifically devoted to project management or process design; however, we chose to go with illustration software out of a sense that our project would be more about clear visual explanation of a process rather than (say) computation of a critical path.

4. Analysis

This is where we apply a programmer's discipline to the diagram, checking its visual syntax as though it were a body of executable code. We look at:

- *Terminology control.* We look for violations of consistent terminology, as when the same term is used to mean two different things, or two terms are used to mean the same thing. As an example of the first problem, we found that different groups would speak of entering data into or extracting data from "The Database" but they were (unknowingly) referring to different databases. As another example, we found that staff used the term "restricted data" to mean three quite different things, depending on the context. As an example of the second problem, we found that the staff were (collectively) using four very different pieces of jargon to mean the same thing: the record of metadata they were assembling for a data collection.

- *Object management.* We look for undocumented or dangling data, materials, storage; objects that are created but don't appear to be used; objects that appear out of nowhere; and objects that appear to be mysteriously moved, copied, deleted, renamed, or otherwise modified. Such anomalies may merely indicate that the diagram is not yet accurate, and some more discovery is called for; but often they reveal that some aspect of the process is poorly understood by the organization, if not outright broken. At ICPSR we found several examples of data, documents, or database elements that were created and handed off to another group, but then never used.

  Object ownership. We look for orphaned processes, which have no clear owner, and contested processes, over which two or more people or groups claim ownership. At ICPSR we saw numerous examples of this around quality control. Nearly every group involved in the pipeline spoke of the valuable checks that they performed, and of the other checks that they believed were being done by other groups. However, on examination of quality control across the entire pipeline, it turned out that certain checks were actually being done several times (redundantly) and others were not being done at all.

5. Validation

This is where we track a sample of actual work going through the process, and update the diagram accordingly. If the analysis step is analogous to program syntax checking at compile time, this validation step is analogous to program checking at run time. Does the program actually execute? Does an actual piece of work follow one of the paths shown on the diagram, or does it enter uncharted territory?

At ICPSR we followed some data collections through the pipeline. We watched as they were handed off from one group to another; we tracked them with sticky notes on the diagram on the corridor wall; and we kept a detailed online diary to record observations.

We found that although the data collections proceeded through the pipeline mostly as described, there were several points of surprise, which revealed situations that we hadn't previously heard about. We followed up with more investigation and then incorporated additional paths and clarification into the diagram.

6. Redesign

This is where we propose changes to the process we've diagrammed.

In the course of constructing the diagram, we find that many of the conversations about how the process currently works are peppered with stories about how the process doesn't work, might work in the future, or should be made to work.

We capture these plans, suggestions, and wishes; and we incorporate them into the diagram or prepare separate diagrams, so as to inform focused discussions on redesigning the process.

Stability and Change

So does this case study support the colloquium hypothesis that "What we do remains the same, how we do it changes"?

We could debate this as stated, but allow me to respond with an alternative formulation that is less black-and-white:

- ICPSR's collection of activities form a spectrum between "what we do" and "how we do it."
- On one side of the spectrum, there is an essence to "what we do" that has indeed remained the same. ICPSR's core service is still preservation and dissemination of social science data collections; we still ingest data, evaluate it, process it, archive it, and publish it.
- On the other side of the spectrum, there are endless details of "how we do it," and they will change endlessly in response to evolution in the services we offer, the technology we use, the organization we belong to, and the communities we serve.
- There is a grey area in the middle, which can look like "what" or "how" according to one's point of view. For example, our webmaster might think of our activity of publishing data on the web to be what we do; long-time staff in other departments might say that data dissemination is what we do, and web publishing is how we do it currently.

Where is process mapping on that spectrum? I will suggest that one of ICPSR's organizational activities has always been "striving to understand how we do what we do," and that process mapping gives us a new way to do it effectively.

New Skills

The sort of process mapping effort we describe here requires the organization to bring together the process content experts and formal process owners (who may or may not be the same people) with the skill sets of five roles:

- an *Investigator*, who contributes skills in effective interviewing, unbiased observation, careful recording, and logical deduction;
- a *Process Designer*, who contributes skills in understanding the logic of complex procedural sequences in a technical organization and making them more efficient and effective;
- a *Technical Communicator*, who contributes skills in clear writing and logical presentation;
- a *Graphic Designer*, who contributes skills in rendering ideas visually by hand or, these days, using software; and
- a *Web Publisher*, who contributes skills in crafting documents for the web, and putting them in place on a website.

And, as with any project, these people could use the support of:

- a *Sponsor*, who provides the organizational resources for the effort;
- a *Project Manager*, who provides coordination and tracking; and perhaps even
- an *Online Document Librarian*, who organizes and maintains the project's collection of graphic and text documents in an orderly file system.

In practice, you may be able to find people who can fill more than one role and contribute more than one skill set.