New Roles for New Times:

Digital Curation for Preservation

March 2011

Tyler Walters
Katherine Skinner

New Roles for New Times:

Digital Curation for Preservation

March 2011

Tyler Walters, Virginia Tech
Katherine Skinner, Educopia Institute

New Roles for New Times:
Digital Curation for Preservation

March 2011

Report Prepared for the Association of Research Libraries by

Tyler Walters, Virginia Tech, and
Katherine Skinner, Educopia Institute

For more information on the series, and to download the PDF, please visit:
http://www.arl.org/rtl/plan/nrnt/

ISBN 1-59407-862-9
EAN 978-1-59407-862-0

Published by the
Association of Research Libraries
Washington, DC 20036
www.arl.org

Contents

About the Authors

Tyler Walters

Tyler Walters is the Dean of University Libraries, Virginia Tech. From 2002 through early 2011, Walters was the Associate Dean of the Library and Information Center, Georgia Institute of Technology. He was a 2008–2010 Fellow in the Association of Research Libraries' Research Libraries Leadership Fellows program. Walters has been involved in many digital initiatives, including Georgia Tech's repository, publishing, and scholarly communication services and led the IMLS-funded Georgia Knowledge Repository, a statewide repository service. Walters is also a founding Board member of the Educopia Institute and Steering Committee member of the MetaArchive Cooperative. He serves on many professional bodies such as the Steering Committee for the International Conference on Open Repositories (http://www.openrepositories.org/), the Interim Governing Board for the Unified Digital Formats Registry (http://www.udfr.org/), the Editorial Board of the *International Journal of Digital Curation* (http://www.ijdc.net/index.php/ijdc), and the Advisory Board for the Digital Information Management program, University of Arizona (http://digin.arizona.edu). He teaches graduate LIS courses for the University of Arizona and for San Jose State University.
Walters has been involved in numerous Federal grants regarding digital repositories, publishing, and preservation. He has presented at over seventy-five conferences and published over twenty-five articles in journals such as the *American Archivist, D-Lib Magazine, International Journal of Digital Curation, Journal of Digital Information, Library Hi-Tech, Library Trends, New Review of Information Networking, portal: Libraries and the Academy* and is a recipient of the Society of American Archivists' Ernst Posner Award for best article in the *American Archivist*. He holds an MA from North Carolina State University and an MA from the University of Arizona. Walters is currently a PhD candidate in Managerial Leadership for the Information Professions, Simmons College, Graduate School of Library and Information Science.

Katherine Skinner

Dr. Katherine Skinner is the Executive Director of the Educopia Institute (http://www.educopia.org), a not-for-profit educational organization that acts as a catalyst for collaborative approaches to the production and preservation of scholarship. She also serves as the Program Manager for the MetaArchive Cooperative (http://metaarchive.org), a distributed digital preservation solution for cultural memory organizations. She was previously the digital projects librarian at Emory University, where she served as co-PI on numerous projects in digital scholarship, access, and preservation arenas.

Skinner received her Ph.D. from Emory University, and has co-edited two books, *The Guide to Distributed Digital Preservation* (Educopia Institute: 2010) and *Strategies for Sustaining Digital Libraries* (Emory University: 2008), and has authored numerous articles, including "The MetaArchive Cooperative: A Collaborative Approach to Distributed Digital Preservation" (*Library Trends*). She is one of the founders and the former Managing Editor of the *Southern Spaces* Internet journal and scholarly forum. She has served as a faculty member specializing in the topic of digital preservation for numerous workshops, including the "Stewardship of Digital Assets" series (NEDCC, 2007–2009) and the "Staying on TRAC" series for digital collaboratives (Lyrasis, 2009–2011). Skinner also regularly consults with groups that are undertaking digital preservation planning, policy creation, or implementation for their cultural memory-oriented collections.

Executive Summary

In the 21st century, ARL libraries are increasingly exploring and adopting a range of new roles in serving research institutions, researchers, scholars, and students, making the time ripe for ARL to organize a new report cluster focusing on key new roles. The New Roles for New Times series will identify and delineate emerging roles and present research on early experiences among member libraries in developing the roles and delivering services. Each report will describe an emerging role, articulating the audience affected by the new role and the benefits various constituencies experience as a result of the new role. The reports will highlight existing work, report authors' findings, and offer analysis of trends, best practices, and key issues. The New Roles for New Times report, "Digital Curation for Preservation," explores how research libraries are attempting to add value in the chain of events that produce new research knowledge and information.

Digital curation refers to the actions people take to maintain and add value to digital information over its lifecycle, including the processes used when creating digital content. Digital preservation focuses on the "series of managed activities necessary to ensure continued access to digital materials for as long as necessary." In this report, we highlight the intersection of these actions, specifically focusing on how digital curation must facilitate the preservation of our shared digital memory.

We suggest how research libraries need to be repositioned as vibrant knowledge branches that reach throughout their campuses to provide curatorial guidance and expertise for digital content, wherever it may be created and maintained. We argue that libraries can no longer expect that researchers and scholars will come to them for advice and assistance; libraries must instead find new ways to reach them wherever they may be. Research and learning activities are increasingly intra- and inter-institutional, collaborative, interdisciplinary, international, and virtual. We show how the library must adjust its service offerings to this new landscape in order to remain viable.

In the process, we document what we believe is a promising set of emerging roles that libraries currently are carving out in the digital arena. We also highlight and discuss the potential implications of relatively new trends within the research library community, including the outsourcing of services that research libraries have historically provided for their campuses. Finally, we put forward a set of collaborative case studies in the digital curation realm and consider the positive impact of such engagement between research libraries to achieve shared goals.

We assert that the strongest future for research libraries is one in which multi-institutional collaborations achieve evolvable cyberinfrastructures and services for digital curation. The alternative, a "go it alone" strategy, will only lead to dangerous isolation for practitioners, yielding idiosyncratic, expensive, and ultimately unsustainable infrastructures. The report gives readers a thorough appreciation of the emerging practice of digital curation for preservation and how research libraries are fostering curatorial practices in order to ensure that their parent institutions continue to realize their core mission of creating, disseminating, and preserving knowledge.

New Roles for Research Libraries

Today's research libraries are focusing less on the public and technical services paradigm of old (front of the house and back of the house, so to speak), and more on building the trio of strong infrastructures, content,

and services. In this new trio, infrastructure includes facilities, technologies, and the human expertise applied to the organization. Content refers to all the information resources the library makes accessible, including its growing campus-born digital collections as well as its licensed or purchased electronic resources. Services include traditional information services and more emergent ones in the virtual realm, such as information production, access and dissemination, and long-term curation and preservation (e.g., the "embedded research librarian" role). Each of these three components of the library—infrastructure, content, and services—are directly impacted by the paradigm shift research libraries are experiencing at both cultural and institutional levels.

Cyberinfrastructure and Collaboration

Cyberinfrastructure encompasses the overall research environment that libraries are aiming to create. It will enable researchers to communicate efficiently, promoting their ability to create and disseminate their work without having to invent the tools of creation and dissemination every time they undertake a new project. Libraries cannot hope to build a stable cyberinfrastructure unless they work together in collaborative units, investing in community-generated solutions rather than insisting on building individualized workflows and systems. Working in conjunction with other institutions promises to deliver stronger infrastructures and collections that can work in interoperable manners and that do not depend unduly on a singular vision.

Content Acquisition and Hosting

Research libraries are extending their usefulness and presence through offering comprehensive and dependable digital curation and preservation services in support of the intellectual content production activities taking place on their campuses. They are achieving this goal through hosting a broad range of content, including digitized collections, licensed content, acquisitions (e.g., web archiving and manuscript collections), research data and other primary source materials generated on campus, e-prints of the campus's intellectual output (e.g., publications, ETDs), instructional materials and other multimedia assets, and digitally captured lecture series, symposia, and other campus events.

Digital Publishing, Curation, and Preservation Services

The new roles in content acquisition and hosting are complemented by new roles in content production, as well as curatorial and preservation activities. The significance of creating strong partnerships between the library and the academic community cannot be overstated. Libraries are well positioned to provide e-publishing options for their campuses, including in partnership with the surviving university presses. Experiments with e-publishing through libraries have been highly successful to date, particularly in the areas of open access journal production, e-prints publication, and scholarly editions published as digital humanities/social sciences resources.

New Roles for Research Librarians

There are at least seven distinct roles for librarians emerging in response to the changing information management needs of today's researchers, scholars, teachers, and students.

Acquisitions and Rights Advisors

In addition to establishing that content is of sufficient quality, relevance, and overall balance to justify its collection and maintenance, research libraries now must grapple with new criteria such as the stability, viability, and authenticity of the digital files at the point of acquisition. With the assistance of legal counsel, librarians and archivists are creating and reviewing policies regarding digital acquisition and donations to ensure that sufficient rights are granted to the library that it can properly curate and preserve that digital

content. Increasingly, research libraries are employing rights management specialists themselves rather than relying on the academic campus's general counsel, in part because the increasing intellectual property issues require information management expertise as well as legal expertise.

Teachers/Instructional Partners in Learning Spaces
This role includes physical and virtual components, both of which have been directly impacted by the growing information management needs of digital scientific, social science, and cultural heritage materials. Librarians are becoming increasingly involved in campus engagement overall, including forging instructional design and classroom partnerships, providing reference/help services in virtual, as well as physical, environments, and offering continuing education-oriented outreach to the local community.

Observers/Anthropologists of Information Users and Producers
Librarians are spending time "living among the natives" on their campuses and studying their information production and consumption habits. If research libraries hope to meet the future needs of their campuses, they must know their constituents—from administrators and staff to professors and students. As technology continues to develop at a rapid pace, the needs of research institutions will likewise quickly shift.

Systems Builders
This role is about designing and architecting physical spaces such as information commons; virtual spaces, such as websites, digital repositories, and other information systems; and discrete system-building activities such as licensing, metadata creation, and copyright research. It also includes providing scholars with new mechanisms for publishing their manuscripts and articles as e-books and e-journals, as well as more specialized digital resources.

Content Producers and Disseminators
Librarians' historical focus on content curation—including selection, management, and preservation—is a defining component of research libraries' ongoing work. A new role is emerging—that of digital curation, or attention to the lifecycle management of digital objects and collections. The role has two distinct, but related, parts: access and preservation. In the access-oriented digital realm, libraries are managing more information than ever before and ultimately assist users in reaching the content they seek. Digital curation also responds to the long-term needs of digital content, which differs significantly from its physical predecessors. It is now evolving in response to growing concerns about the viability, authenticity, and sustainability of digital content.

Organizational Designers
Organizational design can provide the agility and experimental atmosphere necessary to transform an institutional structure to fit its new terrain. Experimentation with new organizational forms, including distributed and decentralized inter-institutional models of cooperation, is imperative to maintain successful, transitioning research libraries. The librarians who tackle this experimentation via organizational design methodologies bear the burden of helping existing staff members find ways to adapt and learn new skills. They also are working to redefine how research libraries organize their division of labor in ways that better fit the digital and physical landscapes they balance.

Collaborative Network Creators and Participants
Librarians are forming collaborative entities that reach well beyond the grant-project timeframes of old. With intentionality, institutions have worked to foster long-term, collaborative, sustainable practices to support some of their common needs, including technology development, repository management, and preservation.

Librarians are beginning to carefully coordinate both intra- and inter-institutional alliances and maintain clear documentation about the roles and responsibilities undertaken by each division or institution in multi-pronged efforts. They are also learning how to provide the glue that can help to forge healthy collaborative relationships in multiple settings.

Collaborative Strategies for Digital Curation and Preservation

The report briefly describes a cluster of emerging digital preservation programs that ARL libraries are creating as well as participating in. These hold much promise as path-breakers in the field. They also hint at the new landscape that is forming, including the potentially large role that collaborative work between institutions might play in preservation. The profiled institutions include: the National Digital Stewardship Alliance, Chronopolis, Florida Digital Archive, MetaArchive Cooperative, Council of Prairie and Pacific University Libraries (COPPUL) Private LOCKSS Network, Alabama Digital Preservation Network (ADPNet), HathiTrust, DuraSpace, Data-PASS, and the University of California Curation Center, California Digital Library.

Recommendations

Wherever possible, collaborative or community-based approaches to digital curation are likely to be more effective and sustainable. However, effective use of collaborative strategies and community-managed resources requires local investment and capacity building. The following recommendations reflect this reality.

- **Stop waiting and start proactive engagement locally**: This should be done through relationship-building with the Office of Sponsored Programs, the Provost, the Vice President for Research, the Deans and department chairs, through other grant connections, data centers, and individual professors and researchers in your institution. Faculty partnerships are a significant driver in digital curation initiatives as are policy partnerships with campus offices.

- **Stake a claim in the production cycle**: Consider new services in creating multimedia and other digital assets by hosting e-publishing services (e-journals, e-books, e-conference proceedings, etc.) and by hosting and curating digital archives, datasets, digital art objects and websites, etc. The important step is to bring these services in-house and seize the moments of opportunity as they arise.

- **Start retraining and repurposing staff**: The library needs to think of digital curation as a core function of the library and to invest financial and other resources into it accordingly. Seek in-depth, long-term training programs for your interested staff. Bring experts to your library. Maintain the daily conversation that your library is and will be engaging in digital curation services.

- **Be a doer, not a broker, wherever possible**: There is no need to hand over our future to external groups; research libraries have adequately demonstrated that it is possible to collaborate in managing digital content and maintaining staff and technology infrastructures in an economically viable way.

- **Consider digital curation collaborations**: Reach out to partners and together, pick a project or a meaningful aspect of technology infrastructure, and begin building a long-term relationship.

- **Actualize collaborative engagement**: Work collaboratively and steadily with other selected institutions over time to build a sustainable cyberinfrastructure. Early experiences and trends indicate that multi-

institutional collaborative approaches provide a successful organizational context that is necessary to meet this large and ever-shifting challenge.

Research libraries have much to gain if they behave as active players in the development and management of research cyberinfrastructure across the arts, humanities, social sciences, sciences, and engineering fields. If they embrace their emergent roles as anthropologists of research environments, co-producers and broadcasters of digital content, and builders of systems for research collaboration, communication, and scholarly object management, then research libraries and their employees will successfully transform themselves into highly regarded service entities and partners in the global digital research community of the early twenty-first century.

1

Context

Introduction

As most research libraries are well aware, myriad forms of complex and costly born-digital information objects now proliferate, including a wide variety of content types from e-science, the social sciences, and the digital humanities. Across these disciplinary domains, scholars are producing digital information of intellectual value that includes new forms of scholarship, scientific data, notes, electronic records, arts and new media, multimedia learning objects, user-generated web content, and the products of mass digitization efforts. A new set of roles is emerging that seeks to leverage the transformative nature of digital information, the World Wide Web, and the correspondingly emergent models of cyberlearning and e-research. Will these roles be filled by libraries and librarians? Or, will other entities ultimately provide these services to our campuses?

The *New Roles for New Times* report, "Digital Curation for Preservation," explores what we believe to be at stake for research libraries in the early twenty-first century—namely, our own missions as institutions. Herein, we suggest how research libraries need to be repositioned as vibrant knowledge branches that reach throughout their campuses to provide curatorial guidance and expertise for digital content, wherever it may be created and maintained. We argue that libraries can no longer expect that researchers and scholars will come to them for advice and assistance; libraries must instead find new ways to reach them wherever they may be. Research and learning activities are increasingly intra- and inter-institutional, collaborative, interdisciplinary, international, and virtual. We show how the library must adjust its service offerings to this new landscape in order to remain viable.

In the process, we document what we believe is a promising set of emerging roles that libraries currently are carving out in the digital arena. We also highlight and discuss the potential implications of relatively new trends within the research library community, including the outsourcing of services that research libraries have historically provided for their campuses. Finally, we put forward a set of collaborative case studies in the digital curation realm and consider the positive impact of such engagement between research libraries to achieve shared goals. We assert that the strongest future for research libraries is one in which multi-institutional collaborations achieve evolvable cyberinfrastructures and services for digital curation. The alternative, a "go it alone" strategy, will only lead to dangerous isolation for practitioners, yielding idiosyncratic, expensive, and ultimately unsustainable infrastructures. The report gives readers a thorough appreciation of the emerging practice of digital curation for preservation and how research libraries are fostering curatorial practices in order to ensure that their parent institutions continue to realize their core mission of creating, disseminating, and preserving knowledge.

As research institutions accelerate their pace in generating and acquiring digital content, the emerging subfields of digital curation and digital preservation are rapidly rising in significance. The phrase **digital curation** refers to the actions people take to maintain and add value to digital information over its lifecycle, including the processes used when creating digital content.[1] **Digital preservation** focuses on the "series of managed activities necessary to ensure continued access to digital materials for as long as necessary." [2] These highly complementary concepts explicitly show that curatorial actions must serve the needs of current

and future users. In this report, we highlight the intersection of these actions, specifically focusing on how digital curation must facilitate the preservation of our shared digital memory.

A Critical Moment for Research Libraries

Sweeping, large-scale initiatives that produce increasing masses of digital information resources are forever changing the world around us. The world of libraries (and in particular, research libraries) is changing drastically in a rapidly evolving information landscape of mass digitization, electronic publishing, digital repositories and archives, virtual communities, user-generated content, digital learning media, and digital datasets generated by academic communities, corporate organizations, and everyday citizens.

The research library is experiencing a critical moment, one that has serious consequences for the future of the institution. As we have learned in the early digital decades, the curatorial practices established over centuries for managing analog and physical materials do not translate fluidly into the management and care of digital content. The work to integrate new services into libraries, usually without adequate budget expansion, has resulted in an important initial set of experiments that lay the groundwork for different directions our field can pursue. Today, the research library stands at an important crossroads, and the decisions that it makes about its practices and priorities in the digital realm can be used to establish the research library's central leadership role in digital content management, one consistent with, though different from, its analog history. Alternatively, those decisions ultimately may relegate the research library to a different role, that of a middleman that brokers digital content services for the parent institution and the scholarly communities that it serves.

In other industries, from banking to journalism, we have already seen how the shift from print to digital production has allowed specifically for the emergence of new fields and for the development of new practices in established fields, sometimes with unanticipated effects. In this report, we raise questions about the roles that leaders of research libraries are creating for their institutions and staff today, as well as those that they are outsourcing to other entities, and we analyze the implications of these decisions for the future of our own field. Key to this investigation is understanding that the role the research library has played in the past need not be the role that the research library plays in the future. Indeed, the one thing that can be counted upon is that in the digital realm there is increasing competition from the commercial sector for many of the services that research libraries have historically provided to scholars, researchers, instructors, and students. In this moment of field transformation, we posit that libraries are making choices that will inform their future roles significantly.

Embracing an Embedded Position

In 2007 David Lewis offered a 20-year projection for the evolution of research and academic libraries (2005–2025).[3] Three of Lewis' five strategic paths may speak to a near-future role for librarians as digital curation experts that foster preservation of digital collections in research libraries. Lewis highlights that libraries must "migrate the focus of collections from purchasing materials to curating [digital] content." Another path highlights the migration from print collections to electronic resources, including expansion of research library activities to include curating and preserving emerging open digital collections. Lewis argues, though, that libraries will have to choose deliberately to play such a role with regard to these materials.[4]

The other key strategic path Lewis maps out is to "reposition library and information tools, resources, and expertise so that they are embedded into the teaching, learning, and research enterprises." Lewis posits a new "embedded librarian" role, where librarians make themselves essential components of university research teams and serve their institution by helping scholars and students to create and curate research data and secondary information resources generated and published from the research activity. Embedded librarians connect with the scholars and students wherever they are doing their work to deliver these

services. They enhance the creation and management of content by promoting repository services and virtual communities. They produce digital tools for communication and knowledge sharing. In order to play these roles, though, librarians have to find ways to implant themselves in scholarly practices, getting to know scholars in ways that are rare in today's academic institutions.

Research divisions that need to respond to growing concerns about the persistence of their data (including NSF-funded research ventures) may not want to turn data management over to the library, but they may be willing to work and consult with experts from the library who come to them with credible knowledge, experience, and practices in research data management and curation.

Lewis asserts that all of these changes will occur with little to no growth in library budgets. Indeed, most research library budgets are currently shrinking rather than growing. The challenge libraries face is that of using the scarce resources at their disposal to experiment with new, innovative roles within their parent institutions. Such experimentation must seek to achieve a change in traditional practices and in the conventional role of the library by capitalizing on the opportunities that our current paradigm shift provides for new ways of producing and curating both primary research materials and scholarship.

The Power and Pitfalls of Collaborative Strategies

In the digital realm, libraries will still collect, organize, preserve, and disseminate a wide variety of information and data resources, but what additional services they must provide and how they will accomplish the tasks that undergird these core missions remain key questions. It is widely understood that library digital infrastructures are not likely to operate in a solo manner, institution by institution, but instead will require collaborative efforts in order to build a broader-based cyberinfrastructure.[5]

Early examples of such strategies include community-based collaborations between libraries, library partnerships with third-party service providers or corporations, and outsourcing. None of these options are new to the digital arena; historically, most libraries have used some mix of individualized, partnership-driven, and outsourced activities to meet the needs of their audiences. What is different, with regard to digital curation, is that the landscape itself has shifted with the rise of digital technologies. As demonstrated by such (relatively) new and innovative companies as Google, the field of information management is no longer the uncontested terrain of libraries and other not-for-profit groups. Other entities are actively vying for roles—and market shares—in creating, managing, distributing, and preserving content in the digital arena. We also see successful experimentation by new organizational types (e.g., the decentralized, community-led models provided by such companies/initiatives as the Apache Foundation and the Wikimedia Foundation) heralding significant changes in the ways scholars and researchers do their work.

The question is no longer whether, but rather how to collaborate. Numerous reports, particularly those on cyberinfrastructure-building, have highlighted collaborative approaches as key to the success of our libraries as we seek economically viable ways to provide a new suite of services to our campuses and their constituents. However, collaboration does not always prove to be cost effective, scalable, or sustainable. This is largely because collaboration by definition requires compromise. We must carefully consider the implications of the compromises we make as we engage in collaborative activities, both as individual libraries and as a field of librarianship.

Perhaps the most important factor we must consider as we engage in collaborative activities is whether an initiative is library driven or third-party, service-provider driven. In library-driven initiatives, libraries retain authority and ownership over the content, services, and infrastructure they jointly create and maintain. Libraries may work closely with other libraries, corporations, and not-for-profit organizations to achieve their goals in these initiatives, but only library leadership in this process ensures that library interests control the digital assets and the cyberinfrastructure that supports the lifecycle management processes.

In third-party, service-provider driven initiatives, libraries often cede ownership of content and infrastructure to external agents. We have seen quite clearly through the so-called "scholarly journal crisis" of the past two decades that the motivations of external agents often fail to align well with our own. When money-making becomes a primary driver, the economic gains that we initially anticipate via collaborative activities may become substantial, field-wide economic losses, as controlling agents from beyond the campus increase prices and create packages that are money- rather than access-and-preservation-driven in nature.

The role of collaboration in digital asset management is essential. Libraries' implementation of these collaborative activities is still in early, formative stages. New models are emerging, some of which will serve the research community well; others may ultimately compromise libraries' ability to serve their patrons. Will the library continue to serve scholars and researchers and provide a home—usually a publicly accessible one—to society's cultural and scientific heritage? Or will the role it has played as an established infrastructure within the information management field in the pre-digital era be met by other types of entities in a more privatized fashion as we progress in the digital age? The decisions made today by the research library community will determine its future role as the field's digital practices solidify.

Current Trends in Research Institutions

Key trends in the research institution today include administrative shifts toward a more business-like model for running the institution; technical transformations in research, teaching, and learning environments; and new, emergent approaches that are appearing in the growing information, search, and publishing industries.

Among the many trends and external factors impacting the development of research libraries today are the following:

- **Higher Education Administration:** The university's desire to measure its return on investment continues to grow. As a component of the campus that does not explicitly generate revenue, but rather depends on institutional funding to provide value-added services that foster excellence in scholarship, research, and teaching, libraries and librarians are often challenged to create metric-based measurements that accurately reflect the contributions they make to the campus.

- **Research, Teaching, and Learning:** In most academic landscapes, two distinct groups support these activities: the library and the IT division. Increasingly, these groups are debating the best ways to preserve content over long periods of time. These debates have implications for the future of academic institutions' content curation responsibilities—will those reside with the library, with IT, in collaborative practices between these campus units, or perhaps with an outsourced, commercial third party?

- **Information / Search / Publishing Industries:** Initiatives such as JSTOR, Elsevier journal packages, the Google Books project, various Microsoft initiatives, the Apple i-products, Amazon's array of web-based services including its "Kindle" eBook reader, and the influence of Internet service providers, publishers, and their mergers, are all helping to shape the information landscape. All of these entities see the academic market as an important component of their own revenue streams, and all of these new models have implications for the future landscape of digital content management.

There are also many points of divergence and difference in universities, including cultural differences between the various academic disciplines; the tenured faculty, pre-tenured faculty, and students; and each of these groups' varied approaches to conducting research and producing scholarship. What roles may libraries play as cultural and knowledge custodians in this new and highly complex information ecosystem?

Research libraries, through digital curation and preservation services, can contribute to these core functions by sustaining access to the data, information, and knowledge resources that researchers in these institutions create and use as they conduct their research, disseminate their findings, and teach their students.

Implications of the New Information Ecosystem for Research Institutions

One outgrowth of this digital transition is captured by the concept of the Network as the Platform (NasP), which involves all manner of learning, communicating, collaborating, researching, and information-producing activities.[6] Intellectual content is increasingly digital, and users learn, communicate, research, and publish via tools such as Facebook or open source software such as Drupal, and through more targeted, open academic technologies such as HUBzero, DSpace, Fedora, WordPress/ScholarPress, and Zotero. Users want to discover and track all of their information resources from the digital spaces where they are already working. The curatorial implication is that digital information resources of interest must remain exchangeable through differing technologies. Yet they must also maintain their true character and form in order to be recognizably reliable and authentic. Preserving these qualities in the information resource is a core purpose of digital curation for preservation.

The user now functions as a content creator and not just a consumer. They generate and disseminate new intellectual content based on their interactions with other learners and information. They increasingly gravitate towards tools and social networks related to their research and learning, which enable and enhance these experiences. As part of this digital growth spurt, research libraries need to learn how to leverage these virtual research and learning communities, as well as their digital tools. They must consider how to curate and preserve not only the core data and information resources researchers produce, but also the communications shared between researchers through such systems. This is another challenge for the research library engaged in digital curation for preservation.

Embracing the Network as the Platform, libraries create services, release them early, observe their users, and reiterate the process quickly to mature and improve new network-based learning and interaction services. They use the NasP Model for integrating key information service strategies, processes, and goals. For instance, libraries are increasingly inviting users to participate in the enrichment of the information resources they use (e.g., tagging important content to help others find it, correcting and elaborating on metadata when they are doing their research).

Figure 1 offers an illustration of the evolving relationships between libraries and users in the NasP environment:

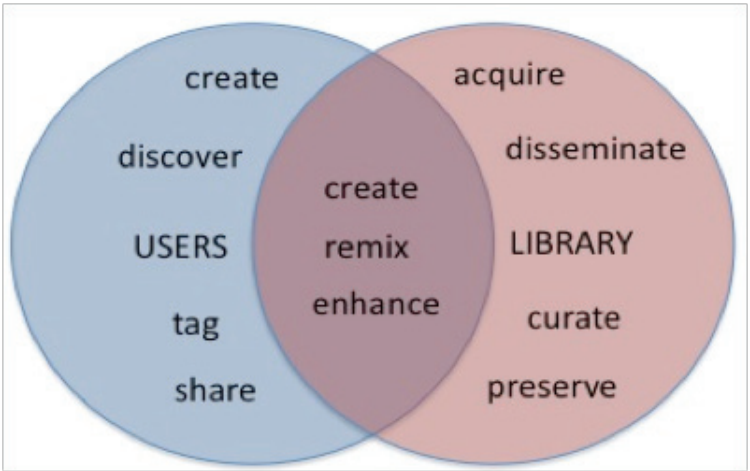

Figure 1: Library/User Roles of the Early Twenty-First Century

Currently, research library functions overlap significantly with user functions. The user may, of course, create, discover, tag, and share content without the library's input or participation. Libraries, however, also build tools and services that enable users to create, remix, and enhance content in partnership *with* the library. Such shared work enables libraries to help users create and work with objects in sustainable ways. Libraries also continue to carry sole responsibility for some functions, including acquisition, dissemination, curation, and preservation of content.

Aligning the work of research libraries with that of their patrons enhances both the patrons' experiences and the libraries' offerings. Moving this work to the center of the service offerings of libraries is an essential next step in the transition toward an effective digital presence. Some of the best work in this area to date has been pioneered through grant-sponsored initiatives, often as shared projects between professors and libraries.[7] In support of this shift to an information-producer focus, research libraries must provide ongoing, dependable services to facilitate and ensure the usability of the researchers' digital works for the long term.

Emerging Conceptions of Digital Curation

Data curation, digital curation, and digital preservation are terms that are variously used to address what are still very much emergent activities and roles. These three terms often correspond to different disciplinary contexts: data curation is applied most often in science, engineering, and social science fields; digital curation is used more frequently to describe digital humanities and arts environments; and digital preservation usually appears in library activities. It is very understandable that different groups of scholars and their library partners have evolved some terminological differences, but these can obscure the fundamental unity in libraries' roles and services relating to digital curation and preservation.

So how do the concepts of digital curation translate across the three main disciplinary tracks of the university—the sciences, social sciences, and arts and humanities—especially as these very tracks arguably show signs of integration through interdisciplinary work? And how might this influence the positioning of the library as a curatorial agent within the overall campus setting? We briefly discuss these issues here; for an in-depth analysis of these issues, please see the *Appendices* of this report.

Differences of Perspective

There continue to be significant differences between the three main research domains of academia that research libraries cannot afford to ignore or gloss over as they reach out within these communities. Consider the impact such factors as longstanding disciplinary divides, physical segregation via campus buildings and quadrants, and distinct, track-oriented funding sources have on these three groups (e.g., NSF, NEH, SSHRC, or NSERC). Most professors know only a small cadre of other professors on campus from beyond their own discipline, and as a result they may not know that they increasingly share common issues, particularly with regard to the digital resources they are creating, maintaining, and using on a regular basis. It is little wonder, then, that they are using different terminology to describe their content management needs. A danger here is that each of these academic tracks could unknowingly pursue separate solutions. Such a silo-based approach is neither cost effective nor as sustainable as a more unified, campus-wide, and even multi-institutional approach.

The curatorial and asset management needs encountered by researchers and professors within each of these disciplinary tracks—whether termed data curation, digital curation, or digital preservation—share more similarities than they do differences. They are often driven by the same basic software frameworks, including databases, mapping tools, and display environments that can handle multiple file types. Most are in need of metadata, data normalization, migration of outmoded formats, and stable and sustainable access and preservation environments. As libraries, we have the potential to cut across discipline-specific divides,

including terminology, by providing a suite of services that appeal to the scholar-creators in each track, but that do not require different discipline-based approaches to long-term maintenance.

Libraries can proactively shape their campuswide digital curatorial landscape by actively engaging with content creators and targeting them where they work, which is often within one of these three main sectors of academia. As we discuss in the next chapter, this necessitates reaching out in new ways across the campus setting, as scholars and researchers do not come to the library for digital content management expertise. For example, libraries currently may capitalize on new requirements by grant agencies regarding data curation and preservation practices by establishing and marketing these services to campus constituents and through developing new relationships with campus divisions that support grant applications (Offices of Sponsored Research and specific positions, such as deans or vice presidents of research, whose offices assist professors with proposals on a regular basis).

We must again note that the overall digital scholarship field is in an early stage of its development. In this still-emerging digital landscape, there are not yet solid centers of production for scholarly works. This niche will be filled, whether by campus-based entities (e.g., university press-type structures or digital centers or libraries) or by external entities (e.g., commercial publishers). Libraries have a brief opportunity to demonstrate their value in such production efforts. If they miss this moment, they risk ceding their digital curation role to other campus or non-campus units and becoming irrelevant to this process. But if research libraries assertively take on this role, they hold the promise to become vibrant think tanks and digital production and management zones, thus building on their traditional role as a content caretaker.

Endnotes

1 See: "What is Digital Curation?" http://www.dcc.ac.uk/digital-curation/what-digital-curation.

2 See definition of digital preservation: http://www.dpconline.org/advice/preservationhandbook/introduction/definitions-and-concepts.

3 David W. Lewis, "A Strategy for Academic Libraries in the First Quarter of the 21st Century," *College & Research Libraries* 68, no. 5 (September 2007): 418–34.

4 Lewis, pp. 427–28.

5 See, for example, the *NSF-based Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure* (http://www.nsf.gov/od/oci/reports/toc.jsp) and the parallel ACLS-based *Our Cultural Commonwealth* report (http://www.acls.org/programs/Default.aspx?id=644), which arguably set the stage for the many authored pieces on "cyberinfrastructure" across the sciences, social sciences, and the humanities this decade.

6 See Tom O'Reilly's writings for more about the Network As Platform (NasP) concept: http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html. See also where Cisco Systems has noticed this concept: http://newsroom.cisco.com/dlls/2008/prod_042208b.html.

7 See for example the Voyages and Origins projects undertaken by Dr. David Eltis of Emory University, which provides ways for both researchers and the general public to contribute to these slave-trade knowledge resources that are then maintained by the library (http://slavevoyages.org and http://www.african-origins.org/). As another extremely successful example, see the Center for History and New Media's (George

Mason) suite of services, which include Zotero for citation, Omeka for publishing exhibits, ScholarPress for building on WordPress blogging software, and emerging text mining offerings. Also note that while this work has been connected to the library and librarians, it has been driven by scholars.

2

Research Libraries and Librarians in the New Information Ecosystem

Research libraries and staff are studying the emergent digital environment and responding to it by generating new library organizational paradigms and service roles, including digital preservation. As trends begin to form, research libraries must evaluate the presence that they wish to have in their parent campus settings and respond accordingly. This is a crucial moment for demonstrating their ability to serve their campuses in relevant digital service areas, including some areas (e.g., scholarly publishing) that have been hosted traditionally by other groups. If desired, research libraries can transform their roles in powerful ways, which may have a tremendous impact on the way scholarship is facilitated, disseminated, and preserved.

Doing so requires libraries to strategize, both individually and as a community. As Margaret Hodge has written, libraries that do not embrace change in this critical time are "sleepwalking into the era of the iPhone, the e-book and the Xbox without a strategy." Libraries that have no strategy, Hodge further argues, run the risk of turning libraries into "a curiosity of history, like telex machines or typewriters." [1] Although not every research library has composed its strategy for transformation to meet the demands of the digital era, many are experimenting with a range of new services.

The following set of new roles that research libraries have been cultivating may serve as markers of the community's recent innovations and leadership. They also may provide an early set of strategic directions in which to grow with intentionality, both as a community and as professionals.

New Roles for Research Libraries

Historically, research libraries have organized their work in terms of public services, such as instruction, reference, and collection selection for the user and technical services, such as acquisition, cataloging, processing, and computer systems activities for use in providing access to the collection. However, today's research libraries are focusing less on the public and technical services paradigm of old (front of the house and back of the house, so to speak), and more on building the trio of strong infrastructures, content, and services. In this new trio, infrastructure includes facilities, technologies, and the human expertise applied to the organization. Content refers to all the information resources the library makes accessible, including its growing campus-born digital collections as well as its licensed or purchased electronic resources. The services include traditional information services and more emergent ones in the virtual realm, such as information production, access and dissemination, and long-term curation and preservation (e.g., the "embedded research librarian" role). Each of these three components of the library—infrastructure, content, and services—are directly impacted by the paradigm shift research libraries are experiencing at both cultural and institutional levels.

The success equation for research libraries with regards to digital curation for preservation becomes one of ensuring access to important scholarly, research, creative, and historically valuable digital data and information resources over time and across new technologies. The emerging focus on the triad of infrastructure, content, and services helps to concretize the transformation from bricks-and-mortar institutions to a "clicks-and-mortar" environment.

Research libraries must, however, work together as a community as they continue making this transition. The most challenging issues that they face currently are not solely technical, but socio-technical in nature. New tools and better bandwidth will not solve these challenges; it is more likely that enhanced collaboration across institutions will do so, including through the development of open standards, to which research libraries should adhere. Undergirding all of the new roles that are described in this report, then, is the intensive need for libraries and staff to work in concert to a degree not yet seen. They have been more typically cooperative, and sometimes competitive, than collaborative in their work to date. In the digital realm, success for libraries largely depends upon their ability to establish and sustain new ways of working together. We will discuss some of the most important experiments in collaboration around digital curation for preservation in Chapter 3, where we point to a cluster of collaborative exemplars that are pioneering new models for library partnerships in order to create sustainable preservation networks and frameworks.

Cyberinfrastructure and Collaboration

Undergirding the innovative work research libraries do in creating, disseminating, and preserving digital content is the cyberinfrastructure upon which all of that work relies. Cyberinfrastructure encompasses the overall research environment that we are aiming to create. It will enable researchers to communicate efficiently, promoting their ability to create and disseminate their work without having to invent the tools of creation and dissemination every time they undertake a new project. As the American Council of Learned Societies (ACLS) Commission on Cyberinfrastructure for the Humanities and Social Sciences has warned, "The infrastructure of scholarship was built over centuries with the active participation of scholars. Cyberinfrastructure will be built more quickly, and so it is especially important to have broad scholarly participation in its construction: after it is built, it will be much harder to shift, alter, or improve its foundations."[2]

Fran Berman, formerly the director of the San Diego Supercomputer Center, and now the vice president for research at the Rensselaer Polytechnic Institute, defines **cyberinfrastructure** as:

> "…the coordinated aggregate of software, hardware and other technologies, as well as human expertise, required to support current and future discoveries in science and engineering. The challenge of Cyberinfrastructure is to integrate relevant and often disparate resources to provide a useful, usable, and enabling framework for research and discovery characterized by broad access and "end-to-end" coordination."[3]

Berman places emphasis on not only hardware, software, and networks, but also on human expertise, both in terms of information technology and information management knowledge, as well as the relevant domain science knowledge involved in a particular cyberinfrastructure environment. Collaboration in this realm is key, both for its production and for the intellectual activities that it must support. Dan Atkins of the University of Michigan, and former director of the NSF's Office of Cyberinfrastructure, refers frequently to the need for enhancing scientists' and scholars' ability to participate in team-based academic inquiry across the world. Hence, cyberinfrastructure connotes a robust set of communication technologies as well as a virtual community in which the research-related communication can take place.

As has been often stated, libraries cannot hope to build a stable cyberinfrastructure unless they work together in collaborative units. However, at least to date, calls for such collaboration have been all too often unheeded, leaving them in danger of building many individual infrastructures that are too fragile to survive the waves of administrative change, personnel change, and technological change to which we are all subject. For example, many institutions' systems groups work on in-house solutions that may comply with national and international standards, but are in essence "home grown." Such solutions may work extremely

well for the local environment. However, these one-off productions simply are not maintainable over time. If a key staff member (for example, the person whose vision has shaped the technical development of the home-grown solution) leaves the institution for any reason, what hope does that institution have of maintaining that legacy vision?

As individual institutions, research libraries are safer in the long term if they invest in community-generated solutions rather than insisting on building individualized workflows and systems. Working in conjunction with other institutions forces them to give up some of their autonomy, but it also promises to deliver stronger infrastructures and collections that can work in interoperable manners and that do not depend unduly on a singular vision. By agreeing on a stack framework with unified features (such as schemas), which gives the research library community the best of both worlds, libraries can build and implement modular components that fit unique institutional needs within a more consistent, overarching environment and serve the broader community in a more sustainable manner. As discussed in Chapter 3, approaches emerging from organizations such as Chronopolis, HathiTrust, California Digital Library, and the MetaArchive Cooperative are examples of pioneering, sustainable frameworks.

Content Acquisition and Hosting

Digital content hosting and curation are core components of the library's work that must be supported by the cyberinfrastructure(s) it builds. Research libraries are extending their usefulness and presence through offering comprehensive and dependable digital curation and preservation services in support of the intellectual content production activities taking place on their campuses. They are achieving this goal through hosting a broad range of content, including digitized collections, licensed content, acquisitions (e.g., web archiving and manuscript collections), research data and other primary source materials generated on campus, e-prints of the campus's intellectual output (e.g., publications, ETDs), instructional materials and other multimedia assets, and digitally captured lecture series, symposia, and other campus events. All of these digital resources, born from research and learning programs, pose tough challenges to research libraries, but they also present new opportunities. These resources confront libraries with the need to extend their capabilities and preserve a growing diversity of intellectual output forms that are now integral to the academic knowledge dissemination process.

Libraries are now collecting traditional research library content (e.g., journal publications, research papers, technical reports, working papers, conference papers, lectures, records, personal papers) alongside more informal resources that have intellectual value (e.g., listservs, threaded discussion lists, chat, virtual community sites/collaboration spaces, blogs, wikis, e-mail, etc.). These informal resources provide an important means for libraries to begin to intentionally capture disciplinary debate and development. Research libraries need to study both formal and informal modes of communication and design solutions for capturing and preserving these unique research-oriented resources. Well-structured digital repositories can be a central tool in curating and preserving such resources.

Libraries are beginning to collect other non-traditional intellectual resources today that challenge and perhaps even change the landscape of information service organizations. Learning output by students include multimedia works produced for classes and undergraduate research programs, as well as students' more general presentations, multimedia works, artistic output, journals, newspapers, magazines, and websites. Faculty are increasingly building digital instructional materials that have value both for reuse in subsequent semesters and for future researchers who seek to understand the evolution of such "learning objects" in the early digital era. Research libraries are keeping these items, both from in-person and distance-learning courses, because of the intellectual value of these resources.[4]

Research libraries are also responding to the needs of digital research data curation. These primary resources need to be made available in order to support and advance research, and research libraries are

demonstrating their ability to effectively coordinate the management and dissemination of this content in a streamlined fashion. Librarians also view research data as an extension of scholarly publications; they are the raw points of data that can accompany journal articles and technical papers.

Two of the most promising new areas of growth are those of digital archive acquisition/management and web archiving. Each of these components of future work requires engagement with an unfamiliar range of legal and ethical issues. Each also demands careful attention to issues of selection and long-term management in ways that necessarily transcend the considerations made in the analog realm to date.

In emerging web archiving practices, institutions are increasingly seeking to actively acquire, disseminate, and preserve the content of the Internet for future generations. Obviously, not everything on the web can be (or should be) saved, but some of the content produced and disseminated using the web could aid current and future researchers in understanding the world in which we live. This content—often posted for brief moments of time—is at serious risk, unless libraries make a concerted effort to collect and manage this digital document-based output.

Web archiving initiatives, including the work of the members of the International Internet Preservation Consortium (IIPC), have been initiated often at the national level. But what role shall research libraries play in this emerging content hosting and preservation work? How can libraries and staff collaborate to ensure that they have the proper tools and procedures to ingest content? How will they work through the legal quandaries that arise when acquiring "orphaned" content or content for which no one bears clear responsibility? And what partnerships can be created with subject specialists and scholars to ensure that libraries are gathering only the most important and relevant sites, and that they are doing so in a way that forms coherent collections that can be important legacies for individual institutions in the future? Each of these issues is still unclear at best in this emerging area; however, engagement with these issues and research libraries' work in web archiving at this stage is necessary if they wish to lay the foundation for their continued role as research repositories in the digital age.

In a similar fashion, libraries must begin experimenting with the processes and procedures of acquiring digital manuscript collections, digital art, digital music, and other digital legacy collections that will certainly comprise a large part of their special collections units in the coming decades. Authors, journalists, musicians, and other notable figures for whom research libraries have provided long-term repositories in the recent past are no longer creating their correspondence, art, and other creations in analog form, but rather are producing documents in both digital and analog environments. This raises new issues that research libraries must actively plan for now.

For example, Emory University's acquisition of Sir Salmon Rushdie's personal archive in 2006 included papers, diaries, manuscripts (including unpublished works), and other content both familiar and expected in such research library acquisitions. But this content was acquired in formats that few libraries have yet contended with: a set of old Macintosh computers and out-of-date storage media types. Such acquisitions will soon be commonplace, and the issues research libraries must begin to grapple with as they grow into their new roles as curators of these digital cultural artifacts are deep and varied. For example, what roles (ethical, legal, practical) can digital forensics play in the retrieval of important research content that notables have intentionally or unintentionally deleted from their machines over time? What legal agreements may govern the decisions made regarding such forensics activities? How can libraries care for the wide variety of out-of-date format and media types that their librarians are likely to meet? When is it most advisable to focus on preserving the content and when must the artifact itself be preserved?

The work of research libraries in this area must be geared to meet the needs of future researchers and technology environments that we cannot yet predict. It also must be undertaken even when they cannot achieve the "preservation perfection" to which they might aspire. One of the most difficult dilemmas research libraries face as institutions in transition is the fight against the inclination to wait until they know how to care

thoroughly for such materials before they are willing to acquire them. Such a stance would be to the detriment of this generation's legacy and to the future researchers who will depend upon that legacy to do their work.

As this work begins, one option is for research libraries to begin cultivating relationships now with those individuals for whom those they would like to serve as the future repository. This would enable institutions to begin assessing collections early and undertaking preservation work in collaboration with the content creators prior to the actual acquisition of their legacy materials. Unlike the analog world to which we are accustomed, the digital world in which these notables are creating their personal and public works is fragile and ephemeral in the extreme. If we wait 30 years until they are ready to select a permanent repository to host their collections, it will be too late. However, it is difficult to know how to justify the expense of working with notables with whom we have not yet established acquisition rights. How do we as a community begin to address the communal problem that lies ahead? Establishing mechanisms for ensuring timely curatorial interventions is a necessary new role for the research library to undertake if it hopes to have collections worth managing from this generation.

With the advent of digital content acquisition comes a new set of roles and responsibilities for the research library that is, as yet, both largely unfunded and not yet mandated by their parent institutions. Finding ways to raise awareness of the future benefits the research academy stands to gain from early participation in web archiving and digital archive management efforts is a challenge that libraries must begin to face community-wide.

Digital Publishing, Curation, and Preservation Services

The new roles described above in content acquisition and hosting for the research library are complemented by the new roles that we are forging in content production, as well as curatorial and preservation activities. The significance of this moment of opportunity for creating strong partnerships between the library and the academic community cannot be overstated. In the recent past, scholarly production services were offered through scholarly presses and scholarly societies. The scholarly societies have almost entirely turned their publishing responsibilities over to for-profit companies, such as Elsevier, which now maintain the journal-publishing enterprise in a profit-centric manner. The scholarly presses are largely in decline, in part due to the resulting crisis that led libraries to prioritize journal subscriptions over monograph purchases during the last decade, and also as prices of the former dramatically escalated. Many university presses are struggling for survival. Most have not experimented widely with e-publishing mechanisms.

Libraries are well positioned to provide such e-publishing options for their campuses, including in partnership with the surviving university presses.[5] Experiments with e-publishing through libraries have been highly successful to date, particularly in the areas of open access journal production (e.g., http://SouthernSpaces.org), e-prints publication (e.g., the Cornell-based arXiv), and scholarly editions published as digital humanities/social sciences resources (e.g., http://whitmanarchive.org; http://slavevoyages.org). In many instances, faculty seeking to produce such works through grant funding have been encouraged to work with libraries, and at this point there is a significant body of content that has been successfully co-produced by faculty and librarians. Librarians are well positioned to provide both infrastructure and guidance for faculty to ensure that the works that they produce are created and published in sustainable ways. The danger is that if research libraries do not claim this responsibility, other entities on campus will. We already see evidence of this on many campuses, where academic departments and centers are serving as the nexus of digital production. We also already see evidence that these departments and centers often do not view the library as a suitable curator of their content, but rather are cobbling together new and separate curatorial infrastructures.

In order to ensure success in the digital arena, libraries must build a strong rationale for undertaking new producer-oriented roles on behalf of their campuses, capitalizing on the way that they can exercise digital

library expertise and cyberinfrastructure for long-term hosting of digital content at reasonable costs. They must also forge strong relationships with university administrators to help design ways to incentivize faculty experimentation with these new methods of creating and disseminating scholarship.

As important, libraries must strongly position this production work as one component of the overall life-cycle management that libraries undertake as one of their core services for institutional content. The library's role in curating and preserving this content should serve as both a purpose and a rationale for being involved in the genesis of digital scholarship. In order to make this possible, the library's curatorial and preservation-oriented work must have an on-campus element. If it is entirely outsourced, there is no reason for the library to be involved at all—a research center could likewise outsource this responsibility. The increasingly digital campus will not be likely to expend resources to two entities (the center and the library) to manage digital collections if it can instead expend resources on one (the center, which outsources its curation/preservation needs). In this scenario, libraries become increasingly vulnerable as scholarship turns increasingly digital.

If instead the library embeds the curation and preservation infrastructure and knowledge within its own staffing and digital framework and provides stable, trustworthy, and affordable services to its campus, the library as an institution becomes more secure and influential within its campus setting. As documented in Chapter 3, there are already a variety of models in which research libraries have partnered to create library-led, sustainable data management and preservation services that retain libraries' control and ownership over knowledge, infrastructure, and content.

New Roles for Research Librarians

Just as research libraries are operating in a new information ecosystem, so too are their librarians. There are at least seven distinct roles for librarians emerging in response to the changing information management needs of today's researchers, scholars, teachers, and students. A wide range of job titles is ascribed in different institutional environments to librarians who fit some mix of these new responsibilities, and many times existing positions have grown to encompass digital curation and preservation. Here, we are not attempting to outline a set of new job titles or classifications, but instead to focus on the roles and responsibilities that have recently emerged and are still emerging that are helping to redefine the library landscape of the twenty-first century.

Librarians as Acquisitions and Rights Advisors

Among the steepest challenges, particularly in archives, is to determine what to acquire and preserve for future researchers and establishing what rights research libraries have to the digital content acquired. Special collections librarians and archivists have long been engaged in selection practices. The basic principles of selection may not change dramatically in the digital medium, but the actual practice of selection relies upon different measures. In addition to establishing that content is of sufficient quality, relevance, and overall balance to justify its collection and maintenance, research libraries now must grapple with new criteria such as the stability, viability, and authenticity of the digital files at the point of acquisition. As important, the curators must establish (in conjunction with the content producer, where possible) what digital forensics activities are legally and ethically advisable for such collections. For example, in a traditional manuscript collection, it is not unusual to undertake work to recover erased content (marginalia, for example, or editing notes). Indeed, scholars make careers out of such recuperative work. In extreme cases, entire texts have been salvaged after being overwritten by future authors (see for example the Archimedes Palimpest Project). While the ethical and legal quandaries are minor when excavating such ancient works, they become more perplexing when focusing on digital manuscript collections that are being increasingly acquired. What rights will libraries have to the materials they acquire, and how will their archival units serve as acquisitions and rights managers on the hard questions that digital collections will necessarily

bring? Hard drives often contain traces of deleted works that, while of great importance to the scholarly and historical record, are intentionally erased by their owners. Will manuscript repositories have sufficient rights to capture the content that exists on the media they acquire in the timeframe necessitated by the short timeframe for action that we deal with in the digital realm? Waiting centuries simply is not an option here as it has been in other, print-based recovery work. Archivists are already tackling this new territory in several early digital acquisition projects. With the assistance of legal counsel, they are creating and reviewing policies regarding digital acquisition and donations to ensure that sufficient rights are granted to the library such that it can properly curate and preserve that digital content. Increasingly, research libraries are employing rights management specialists themselves rather than relying on the academic campus's general counsel, in part because the increasing intellectual property issues require information management expertise as well as legal expertise.

Librarians as Teachers/Instructional Partners in Learning Spaces

This role includes physical and virtual components, both of which have been directly impacted by the growing information management needs of digital scientific, social science, and cultural heritage materials. Librarians are becoming increasingly involved in campus engagement overall, including forging instructional design and classroom partnerships, providing reference/help services in virtual, as well as physical, environments, and offering continuing education-oriented outreach to the local community. They are working to redefine the space of the library itself for learning purposes (e.g., learning commons and the "library as place" movement). They are also trying to better insert their expertise into the classroom and the research lab. However, this approach also can contribute to and positively impact the relationship between the research library and the digital content creators.

Arguably, this is the area in which contemporary research libraries are the weakest and most in need of deepening relationships. Subject liaisons have been the primary connections to the academic departments that they serve; increasingly, research libraries are in need of a wider range of connectivity than can be provided by this model. The concept of the "embedded librarian" moves beyond the subject liaison model to forge a digital curation partnership between the creators and stewards of digital content. The embedded librarian focuses not on analog collection building and resource gathering for teachers and students, but rather on assisting scholars and researchers as they create new forms of digital content to help ensure that such content is created with sustainability principles in mind and designed for long-term maintenance. If the output of the research institution is to be viable for future scholars and researchers, its production must be guided by curators that understand and can convey the limitations, fragility, and ephemerality of digital content, as well as the means of creating and contextualizing digital content for long-term usage. This is a role that research institutions desperately need their libraries to play. It is also a role that enhances and stretches continuing roles in assisting professors, researchers, and students with their Geographic Information Systems (GIS) needs and their data mining needs from a user perspective. Helping campus constituents to become responsible *producers* of digital content that is sustainable and well curated is a natural growth area for research libraries engaged in digital curation for preservation.

Librarians as Observers/Anthropologists of Information Users and Producers

Librarians are spending additional time "living among the natives" on their campuses and studying their information production and consumption habits (ideally with just as much emphasis on the production as the use of digital content). If research libraries hope to meet the future needs of their campuses, they must know their constituents—from administrators and staff to professors and students. As technology continues to develop at a rapid pace, the needs of research institutions will likewise quickly shift.

Libraries must continue to assess the compatibility of the services they offer with the needs of their campuses, and to that end they are beginning to more systematically research and anticipate the near-future needs that they will be expected to meet. They can no longer rely on grant-driven research opportunities to build their cyberinfrastructures; they should instead build the case for its development through their work, research, and collaborations with the communities that they serve. One important component of this work is to assess what web-based resources scholars believe might be of highest value to future scholarship and research so that these resources may be harvested and preserved by libraries in a systematic fashion.

Librarians as Systems Builders

This role is not about systems administration in the server room, but rather about designing and architecting physical spaces such as information commons; virtual spaces, such as websites, digital repositories, and other information systems; and discrete system-building activities such as licensing, metadata creation, and copyright research. It also includes providing scholars with new mechanisms for publishing their manuscripts and articles as e-books and e-journals, as well as more specialized digital resources. Researchers are experiencing an important shift in the ways they create and disseminate their scholarship; teachers are likewise experiencing an important shift in the ways that they communicate with their students, both locally and distance-education-based. In both of these capacities, the research environment needs technology partners to help shape the landscape of the future academic commons. Research libraries must continue to work to become a central player in this evolving field.

Much of the work currently handled by librarians as systems builders began as grant-funded ventures. Such research-driven and often pioneering work now must be supported and sustained by systems builders via a mix of self-supported and intra- and inter-institutional partnership-driven work. Systems builders are adopting and creating open standards to help provide a unified development direction for the overall research library field in order to try to avoid the silo effect that has been cultivated in much of research libraries' early technology development work. The willingness of technology leaders to collaborate, rather than building highly individualized systems intended to serve only their campuses, will determine much of the capabilities possessed by research institutions of the digital era. Research libraries cannot afford to create individual mechanisms for every campus—this approach is too specialized and a wasteful use of scarce resources. They are confronting common challenges and need common infrastructure developments to provide the overall cyberinfrastructure that can support the new scholarly communications apparatus. The leaders in the field over the next decade will be the ones who focus on developing standards-based, service-oriented architectures in multi-institutional or community-based practices. They will build modular components in stack-type architectures, thus enabling institutions to choose and/or design components that fit the specifics of their local missions while still building a sustainable whole. The maverick systems developers who work out of sync with such communities increasingly put their own institutions at high risk because they build very siloed infrastructures that cannot be easily maintained without their specific, personalized input and guidance.

Librarians As Content Producers and Disseminators

Historically, librarians and archivists have been looked to as the custodians of physical cultural heritage. This focus on content curation—including selection, management, and preservation—is a defining component of research libraries' ongoing work. A new role within this arena is emerging—that of digital curation, or attention to the lifecycle management of digital objects and collections. The role has two distinct, but related, parts: access and preservation. In the access-oriented digital realm, libraries are managing more information than they ever have before, which requires them to build new mechanisms that help manage the flood or tidal

wave of digital content and ultimately assist users in reaching the content they seek. Digital curation, in part, responds to the demand for better ways of ensuring accessing to content.

Digital curation also responds to the long-term needs of digital content, which differ significantly from its physical predecessors. Similar to the emergence of the subfield of preservation in library science six decades ago—in response to such factors as the brittle books crisis—digital curation is now arising in response to growing concerns about the viability, authenticity, and sustainability of digital content. To ensure the survivability of digital information objects, curatorial action must happen within a much shorter optimal timeframe than that experienced with physical objects, due to the fragility of the bits and the speed with which many technological systems become obsolete. In this new role, librarians are experimenting with new ways of managing the whole lifecycle of digital objects, including creation/acquisition, dissemination, and long-term care.

In the digital curation area, there is a need for staff who can prepare existing digital content for deposit, facilitate its safe exchange between storage media, and ready it for long-term management. Such data wrangling helps to assess and rectify divergent practices in data management (e.g., different file structures, naming conventions, metadata schema deployment) to make the overall infrastructure, as well as individual collections, more sustainable.

Data research scientists, who collaborate with data producers and repository contributors to develop successful strategies for collecting and creating research datasets that are reliable, discoverable, sharable, and sustainable are also emerging. They are data-centric and bear a solid grasp of both digital research data and information science. Providing such data experts in each research division on an academic campus would be costly and would likely result in redundancies and divergent practices across the campus. Centralizing these positions within the library enables the data research scientists to benefit from exposure to a wider range of data and helps to provide a perspective that is united across different campus agents—ultimately resulting in more manageable collections for the long term.

Librarians As Organizational Designers

Librarians in management positions are increasingly called to become organizational learning specialists, or knowledge environment engineers, who design organizational ecosystems. Effective designs help the people who comprise these ecosystems to learn, adapt, and grow as new drivers and motivators appear in the information-service environments. Drawing on fields like organizational development and learning is critical to this type of librarian or information professional. Especially in moments of great change, such as the current transition from a print-driven world to a digital-driven world, organizational design can either provide the agility and experimental atmosphere necessary to transform an institutional structure to fit its new terrain *or* can hold the institution back, ultimately leaving it vulnerable to extinction in the quickly changing environment.

Experimentation with new organizational forms, including distributed and decentralized inter-institutional models of cooperation, is imperative to maintain successful, transitioning research libraries. The librarians who tackle this experimentation via organizational design methodologies bear the burden of helping existing staff members find ways to adapt and learn new skills. They also are working to redefine how research libraries organize their division of labor in ways that better fit the digital and physical landscapes they balance. Some of these designers are library deans. Some are consultants who can bring external perspective to local environments. Some also are managing consortia of institutions and helping them to find ways to capitalize on their shared staffing and resources to accomplish large-scale goals that they cannot approach in isolation.

Librarians As Collaborative Network Creators and Participants

The research library community must act as a community if libraries are to build a sustainable framework for scholarly communication, particularly one that they can manage for themselves. One way that research libraries are doing this is by forming collaborative entities that reach well beyond the grant-project timeframes of old. With intentionality, institutions have worked to foster long-term, collaborative, sustainable practices to support some of their common needs, including technology development (e.g., DSpace, Fedora), repository management (e.g., state-wide infrastructures such as Minnesota Digital Library, the Texas Digital Library, and the GALILEO Knowledge Repository), and preservation (e.g., LOCKSS, MetaArchive Cooperative, HathiTrust, DuraSpace). Creating these collaborative infrastructures in ways that support the extended library community *and* in ways that are dependably strong has required librarians to stretch their skills. In some cases, they have founded non-profit infrastructures that reflect the values and mission of the community, providing a central point of coordination without bulky overhead to maintain (e.g., DuraSpace and the Educopia Institute). In others, they have worked within a single university infrastructure to provide leadership and a base of operations for a joint effort (e.g., HathiTrust, LOCKSS) or used existing state-based resources to provide that central glue for the initiative (FCLA's Florida Digital Archive, Alabama Digital Preservation Network). In every case, both the central unit and the individual institutions that comprise the collaborative network must tend to the network to keep it healthy and viable. Librarians are learning how to establish solid membership agreements and Memorandum of Understanding that can help to legally govern their collaborative work and provide a measure of assurance so that no one institution can easily jeopardize the success of the overall initiative. They are also learning how to obtain the recognition that is still a powerful driver and measure of success within the field, but to do so by making contributions that serve the community, not just their own ends.

Librarians are beginning to carefully coordinate both intra- and inter-institutional alliances and maintain clear documentation about the roles and responsibilities undertaken by each division or institution in multi-pronged efforts. They are also learning how to provide the glue that can help to forge healthy collaborative relationships in multiple settings. These skills are imperative for twenty-first century research libraries. In an environment where outsourcing is becoming ever more prevalent (e.g., Library Systems and Services, a private company which is now the country's fifth-largest library system after taking over more than a dozen library systems), it is becoming clear that research libraries must explore alternatives that run as community-based mechanisms to maintain their role and autonomy.[6]

Endnotes

1 Margaret Hodge, http://www.guardian.co.uk/books/2010/mar/07/future-british-libraries-margaret-hodge.

2 Introduction, http://www.acls.org/programs/Default.aspx?id=644.

3 Examples of cyberinfrastructure in place to support scientific research can be found in reports such as "Cyberinfrastructure Vision for 21st Century Discovery," Washington, D.C.: National Science Foundation (2007), http://www.nsf.gov/pubs/2007/nsf0728/index.jsp.
Other writings describe cyberinfrastructure environments such as the National Virtual Observatory and its relationship with the Johns Hopkins Libraries in curating its astronomical data. See Sayeed Choudhury, et. al., "Digital Data Preservation for Scholarly Publications in Astronomy," *International Journal of Digital Curation* 2, no. 2 (2007): 20-30, http://www.ijdc.net/ijdc/article/download/41/193.

4 See http://ocw.mit.edu/OcwWeb/web/home/home/index.htm.

5 See for example the work done to bridge the scholarly press with the library at Pennsylvania State University in order to address digital publishing concerns (including providing long-term curation for publications).

6 See http://www.nytimes.com/2010/09/27/business/27libraries.html?_r=2&hp.

3

Collaborative Strategies for Digital Curation and Preservation

Digital preservation organizations and services are beginning to form as a key component of the emerging digital library field. Some of these are library-driven operations, in which a central repository is implemented to function locally according to preservation principles (e.g., Library of Congress's Chronicling America collection). Others are community-driven initiatives, in which a group of research libraries partner together to create a preservation infrastructure that enables members to achieve their preservation needs, including distributing copies geographically (e.g., LOCKSS, MetaArchive Cooperative, Chronopolis, DAITSS, HathiTrust). And some are business-driven enterprises, in which third-party service providers beyond the library or campus setting offer preservation services to the library market (e.g., Portico, Ex Libris's Rosetta, OCLC's Digital Archive).

Each of these approaches brings with it both costs and benefits. The central repository model enables an institution to own the whole process of preservation and keep it in house, in effect maintaining control. It may be expensive (both in terms of staffing and technology) and does not readily lend itself to geographical distribution. The community-driven repository or network uses the strength of a distributed community to build preservation solutions that can be geographically diverse. Expertise is built within the institutions themselves, and as such, each constituent library's control over the preservation process is maintained. The community must be bound together carefully, though, in order to ensure that the preservation does not depend unduly on any one member, or its sustainability becomes questionable. The third-party approach enables institutions that cannot technically or organizationally support their own preservation activities to preserve their content through outsourcing. As seen in many other situations, including the infamous "journal crisis," outsourcing is always a market-based activity, and those that participate may eventually find themselves paying higher-than-expected fees. It is also an activity that necessarily limits the control of the contributing institution, including control over both the content itself (which it increasingly chooses to rent instead of own) and the technical steps used to preserve that content, including how that content is migrated from current technical platforms to other, future platforms. Currently, it appears that outsourced services will hold most interest for smaller (non-research) libraries with less robust technical infrastructures and staffing. However, the early success of Portico as a preservation solution for journal content demonstrates that there are already some targeted services of interest to research libraries.

Briefly described below are a cluster of emerging digital preservation programs that ARL libraries are creating as well as participating in. These hold much promise as path-breakers in the field. They also hint at the new landscape that is forming, including the potentially large role that collaborative work between institutions might play in preservation. The first organization described is the National Digital Stewardship Alliance, a new organization founded through the National Digital Information Infrastructure and Preservation Program of the Library of Congress, which provides a network that spans across the growing digital preservation community. Next is a selective overview of some of the most successful preservation programs ARL libraries have helped to foster to date.

The National Digital Stewardship Alliance

Contributed by Michelle Gallinger, NDIIPP, Library of Congress

The mission of the National Digital Stewardship Alliance (NDSA) (http://www.digitalpreservation.gov/ndsa/) is to establish, expand, and promote the capacity to preserve the nation's digital resources for the benefit of present and future generations. This is accomplished through the collaborative effort of the government agencies, educational institutions, non-profit organizations and businesses that make up the 63 members of the newly formed Alliance.

The National Digital Information Infrastructure and Preservation Program (NDIIPP) launched the Alliance in July 2010 as a collaborative national network to support long-term access to digital content. Over the past 10 years, NDIIPP has underscored its commitment to digital preservation by working to strengthen and expand the emerging digital preservation network. The network approach has been effective in leveraging the strengths of a variety of diverse partners, and has proven flexible in the face of technological unpredictability, economic downturn, and the exponential growth of digital materials. The foundation of the National Digital Stewardship Alliance is based on this collective experience. This new initiative leverages best practices and relationships for expanding the stewardship of digital collections by an ever-growing community.

New members will join with other organizations committed to the preservation of the nation's digital heritage. Members share expertise as well as tools and practices to benefit local efforts while contributing to the stewardship of a growing national collection of diverse and significant digital content. Organizations interested in joining the NDSA are encouraged to complete the application form: http://www.digitalpreservation.gov/ndsaform/.

Members of the Alliance will set a strategic digital preservation agenda; shape the plan for advancing digital preservation action in the United States; participate in efforts to develop and implement tools, services, and training; build public awareness about the importance of digital preservation via a national outreach program; and partner with major organizations and refine digital preservation practices with a network of thought leaders. The NDSA is also a mechanism to shape the digital preservation practices of the United States as well as collaborations with international digital preservation efforts.

There is no fee for membership; the Alliance is a volunteer organization. Contribution is made though work. Participants from member institutions of the Alliance work together and make a sustained contribution to digital stewardship through at least one of five working groups.

- **Content**
  The Content Working Group will focus on identifying content already preserved, investigating guidelines for the selection of significant content, discovery of at-risk digital content or collections, and matching orphan content with NDSA partners who will acquire the content, preserve it, and provide access to it.
- **Infrastructure**
  The Infrastructure Working Group will work to build a community of sharing information and best practices about the development and maintenance of tools and systems for the curation, preservation, storage, hosting, migration, or similar activities for the long-term preservation of digital content.
- **Innovation**
  By encouraging and sharing innovative methods of digital preservation practices and technologies, the Innovation Working Group of the National Digital Stewardship Alliance plans to distribute, document, and share emerging practices, while conducting and guiding research and development with engaged partners to find solutions where none exist.

- **Outreach**
  The Outreach Working Group of the National Digital Stewardship Alliance will focus on building relationships with stakeholder communities, preparing and sharing digital preservation information resources.
- **Standards**
  The Standards and Practices Working Group will work to facilitate a community-wide understanding of the role and benefit of standards in digital preservation and how to use them effectively to ensure durable and usable collections. The Group will also develop, recommend, promote, and disseminate information about effective methods for selecting, organizing, describing, managing, preserving and serving digital content, in collaboration with other individuals and organizations where appropriate.

The organizational management of the NDSA is currently being crafted by its members. Thirty-five organizations volunteered to be part of a temporary organizing committee when they joined the NDSA. The aim of this group is to produce principles of collaboration that create a flexible, lightweight partnership that brings together groups from multiple disciplines and sectors with a focus on digital preservation action. This group has identified the need for a coordinating committee which will bring strategic vision to the Alliance. This strategic vision will be employed by the working groups in planning their projects and executing their work plans. The Working Groups are the seat of action in the Alliance; they identify tasks and develop work plans that are accomplished by participants.

Work plans are developed by members based on their interest and commitment of time. This is one of the ways in which the Alliance is a member-driven organization. This focus on outcomes provides the collaborative environment for members to share expertise and leverage knowledge across disciplines. Individual participation is highly valued and recognized and provides the mechanism for the member organizations to demonstrate commitment to digital preservation and to the shared values of the Alliance.

Chronopolis

Contributed by David Minor, San Diego Supercomputer Center, University of California, San Diego

Established in 2007, Chronopolis (http://chronopolis.sdsc.edu/) is a digital preservation data grid framework developed by the San Diego Supercomputer Center (SDSC) at UC San Diego, the UC San Diego Libraries (UCSDL), and their partners at the National Center for Atmospheric Research (NCAR) in Colorado and the University of Maryland's Institute for Advanced Computer Studies (UMIACS).

A key goal of the Chronopolis framework is to provide cross-domain collection sharing for long-term preservation. Using existing high-speed educational and research networks and mass-scale storage infrastructure investments, the partnership is designed to leverage the data storage capabilities at SDSC, NCAR, and UMIACS to provide a preservation data grid that emphasizes heterogeneous and highly redundant data storage systems.

Specifically, the current partnership calls for each Chronopolis member to operate a grid node containing at least 100 TB of storage capacity for digital collections. For reference, just one terabyte of information would use up all the paper made from about 50,000 trees. The Chronopolis methodology employs a minimum of three geographically distributed copies of the data collections, while enabling curatorial audit reporting and access for preservation clients. The key underlying technology for managing data within Chronopolis is the Integrated Rule-Oriented Data System (iRODS), a preservation middleware software package that allows for robust management of data. The partnership is also developing best practices for the worldwide preservation community for data packaging and transmission among heterogeneous digital archive systems.

Chronopolis has concentrated on building a wide range of content that is not tied to a single community. Currently there are four significant collections housed in Chronopolis. Some significant examples include:

- A complete copy of the data collection from The Inter-university Consortium for Political and Social Research (ICPSR), based at the University of Michigan. Established in 1962, ICPSR is the world's largest archive of digital social science data.
- Data from The North Carolina Geospatial Data Archiving Project, a joint project of the North Carolina State University Libraries and the North Carolina Center for Geographic Information and Analysis. It is focused on collection and preservation of digital geospatial data resources from state and local government agencies in North Carolina.
- Scripps Institution of Oceanography at UC San Diego (SIO), which has one of the largest academic research fleets in the world, with four research vessels and the research platform FLIP. Since 1907, Scripps oceanographic vessels have played a critical role in the exploration of the planet, conducting important research in all the world's oceans. SIO is providing data from several decades of data from its cruises.
- Web-at-Risk, a multi-year effort led by the California Digital Library (CDL) to develop tools that enable librarians and archivists to capture, curate, preserve, and provide access to web-based government and political information. The primary focus of the collection is state and local government information, but may include web documents from federal and international government as well as non-profit sources.

Chronopolis is currently in the midst of moving to "generation two" of its preservation system. This includes increasing the storage capacity at all sites as well as moving to the latest installation of iRODS and other core software. In addition, the Chronopolis team is working toward detailed collaborations with other large preservation initiatives in the state of California and throughout the United States. The hope is that in the near future this will provide an even more robust preservation network with multiple services and ingest points. Finally, Chronopolis is now offering its preservation services as a fee-based service to organizations in need. This service is available for immediate use by institutions who need a mature preservation environment without but do not want to create the infrastructure needed on their own.

Florida Digital Archive

Contributed by Priscilla Caplan, Florida Center for Library Automation

The Florida Digital Archive (FDA) (http://www.fcla.edu/digitalArchive/) is a centralized digital preservation repository for the use of the eleven universities in the Florida state university system. The FDA is run by the Florida Center for Library Automation (FCLA), which provides automated systems and services to the libraries of the state system. When established in 1984, FCLA's main job was to run a shared implementation of the NOTIS library management system. Over the years, the mission expanded to supporting an array of productivity and access applications, consortial licensing of electronic resources, and support for digitized and born-digital materials. In 2001, at the direction of the library deans, FCLA entered the digital preservation arena and began developing the DAITSS preservation repository application.

The Florida Digital Archive went into production in late 2006 with no dedicated operations staff, a huge accumulated backlog of materials for archiving, and a partially completed version of DAITSS (you could put stuff in, but not get anything out). As of August 2010, the FDA has full-time manager and operations technician positions, no processing backlog, and about 250,000 stored packages (63 TB). A project

to re-architect the DAITSS application as a series of web services is nearly complete with production implementation planned for 4Q 2010.

Description of approach to organizational management

The FDA is a "dark archive" with no public access and no functionality beyond long-term preservation. University libraries in the state system all run the Ex Libris Aleph system for library management but beyond that are free to select and implement any applications that meet their needs. The libraries use various applications for their institutional repositories, digital library or digital asset management systems, ETD systems, and so on. Regardless of how a digital resource was created or is made available on campus, if the resource is selected for preservation a copy must be sent to the FDA in a prescribed submission package format. The FDA promises to deliver back to the library on request, at any point in time, a copy that is bit-wise identical to the original resource. If all files comprising the resource are in supported formats, the FDA will also deliver a version guaranteed to be renderable with tools available at the time of the request.

In order to archive materials in the FDA, a library must first negotiate a binding agreement with FCLA. The agreement lays out the responsibilities, liabilities, and warranties of both parties, and is signed by the FCLA director and either the library director or university counsel on behalf of the university board of trustees. Once the Agreement is signed, the library becomes an affiliate of the FDA. The term affiliate is used instead of submitter or customer to emphasize that the FDA and the libraries work in partnership. The library deans are the de facto governing board of the FDA, and responsibility for long-term preservation is shared between the FDA and the affiliates.

In this model of shared responsibility, the affiliate is responsible for the following:

- Selecting content to be archived
- Securing rights to archive and preserve the content
- Describing the content adequately for its own purposes
- Submitting packages in format required by the FDA
- Maintaining local records of what it archived
- Withdrawing content that should no longer be archived
- Requesting disseminations when needed
- Providing access to disseminated content

The FDA is responsible for the following:

- Accounting for every package submitted with an Ingest or Rejection report
- Providing useful counts and reporting information on ingested materials
- Implementing preservation strategies as described in the FDA policy guide
- Preserving original files exactly as submitted, with demonstrated integrity, viability, and authenticity
- Providing a renderable version of all supported formats
- Providing disseminations on request
- Attempting to achieve and maintain certification as a trustworthy repository

Brief description of technical achievements

The DAITSS application that underlies the FDA is locally written software designed to implement the OAIS reference model and perform active preservation strategies based on format transformation. DAITSS attempts to identify and describe all files, and for any file in a supported format will create a normalized or migrated version (or both) when possible and desirable.

Some of the hallmarks of DAITSS are:

- The application does preservation and nothing else and as such must function as a "back end" to other systems for acquisition and user access.
- It depends heavily on well-known standards, including OAIS, METS, and PREMIS.
- Archived content is stored with all of its metadata, although some metadata is also replicated in a database for fast access, so the archival store could be interpreted even without the application.
- All format-based processing, including migration and normalization, is done inside the system, which keeps a rigorous record of digital provenance.
- Format-based processing takes place at the time of ingest and as part of a process called refresh—packages are refreshed before dissemination to ensure that they are fully up to date.

Current initiatives

- FCLA developers are in the process of rewriting DAITSS as a set of RESTful web services (DAITSS 2). Each web service stands alone and can be used in other contexts as well as part of DAITSS. The new architecture makes it easier to modify and test the code, enables rapid integration of external third-party tools, and allows for the flexible allocation of resources. For example, there could be multiple instances of a resource-intensive activity such as virus checking, running on the same or different real or virtual machines.
- When the FDA is successfully migrated to DAITSS 2, the DAITSS 2 code will be made available as open source for any commercial or non-commercial site to implement. FCLA is looking into grant opportunities and partnership possibilities to help expedite the distribution and support of the system.

MetaArchive Cooperative

Contributed by Katherine Skinner, Educopia Institute

The MetaArchive Cooperative (http://metaarchive.org) is an international preservation network comprised of research institutions. Established in 2004 through the National Digital Information Infrastructure and Preservation Program (NDIIPP) of the Library of Congress, the MetaArchive model focuses on sharing responsibility, sharing expertise, and sharing cyberinfrastructure to enable libraries, archives, centers, and museums to accomplish their preservation goals as a distributed community. MetaArchive members bring their collective strength to bear on the preservation challenge, not by outsourcing or centralizing operations, but rather by building knowledge and infrastructure in local institutional environments. The technologies are open source, and the curation and preservation services ensure the long-term accessibility of authentic content. Together, the membership preserves a broad range of digital assets, including ETDs, newspapers, journals, and archival holdings (including video, audio, image, and other media types), as well as digital creations from the digital humanities, social sciences, and sciences (such as datasets, databases, portals, and other resources).

Organizational Model

MetaArchive is a cooperative membership organization with three membership categories: Sustaining Members, Preservation Members, and Collaborative Members. Each member runs a server for the MetaArchive network and prepares its own content for ingest in consultation with the Cooperative's central staff.

- **Sustaining Members** are institutions that make the highest financial and technical commitment to the cooperative. They provide the cooperative's leadership and commit to a pioneering role in the emerging field of distributed digital preservation. Each Sustaining Member has one voting representative on the MetaArchive Steering Committee. They offer input and guidance on future development paths for the MetaArchive, including the creation of new data curation tools and reporting tools.
- **Preservation Members** are institutions that preserve content in the cooperative and support its overall infrastructure through running and maintaining a network server.
- **Collaborative Members** are groups of institutions that run a single, shared, centralized repository and preserve this shared content in the MetaArchive network. They also help to support the cooperative's infrastructure through running and maintaining one of the network's servers.

Finances

MetaArchive is funded through a mixed revenue stream that includes sponsored funding/contracts, consulting revenues, membership dues, and fees for services. The cooperative has enjoyed the support of the National Digital Information Infrastructure and Preservation Program (NDIIPP) of the Library of Congress and the National Historical Publications and Records Commission (NHPRC).

Governance and Staffing

The Cooperative is an intentionally lightweight organization that focuses on building infrastructure within member organizations, not within a central service agency. Most of the work of the cooperative is achieved by member institutions, including SIP preparation, ingest, AIP monitoring, and DIP provision when necessary.

MetaArchive is governed by a steering committee comprised of one voting representative from each sustaining member and one representative each from the preservation and collaborative member categories. Leadership of the steering committee is determined by nomination and simple majority vote by the steering committee. The steering committee directs the activities of the cooperative through weekly phone calls and an annual meeting that is held at a member site.

Three committees, the content committee, preservation committee, and technical committee also guide MetaArchive's work.

- The **Content Committee** is responsible for organizing, developing, and documenting content selection practices and MetaArchive SIP preparation guidelines. The content committee also recommends prioritization of new subject- and genre-based archives for the preservation network (e.g., ETD Archive, Newspaper Archive, Southern Digital Culture Archive).
- The **Preservation Committee** is responsible for researching, developing, documenting, and disseminating policies, procedures, and evaluative means for enhancing the MetaArchive Cooperative's practice of trustworthy distributed digital preservation.
- The **Technical Committee** is responsible for developing and maintaining technical specifications and coming to agreements on hardware, software, and networking protocols; overall server architecture; application development; and software maintenance.

The cooperative is currently staffed by 3.5 positions, a program manager, a collaborative services librarian, a systems administrator, and a software engineer. These positions provide the foundation for the cooperative's ongoing work. The program manager oversees the day-to-day operations of the MetaArchive cooperative. The collaborative services librarian trains new members and assists all members as they manage their network servers, prepare their content for ingest, and monitor their content. The systems

administrator oversees the core network functions, monitors content, and helps to train members as they bring up and monitor their own servers. The software engineer assists all members with their SIP development, ensuring that the SIP conforms to MetaArchive data management guidelines.

The cooperative and its membership believes that local staff need to be actively involved in preservation activities in order to maintain a vibrant and knowledgeable preservation community. The cooperative therefore intentionally depends upon distributed staffing located at each member institution. Each member runs an autonomous server for the network, ensuring that every copy of content is maintained by a different system administrator and that there is therefore no one point of human failure within the network. Each member also works in concert with the central staff to prepare its content for ingest. Most members further contribute to the cooperative through committee assignments. These roles are imperative for the cooperative and its membership, both in practical and philosophical terms.

Technical Model

MetaArchive's technical model is grounded in the principle of distributed digital preservation. The central assertion of the cooperative is that research institutions can and should take responsibility for managing their digital collections, and that such institutions can realize many advantages in collaborative, distributed, long-term preservation strategies.

Historically, the most effective preservation efforts have succeeded through some strategy (intentional or not) of distributing copies of content in secure, distributed locations over time. Many of the threats to obsolescence in the digital arena are the same (natural disasters, intentional attacks, accidental destruction) as those faced in other eras. To be sure, there are additional challenges that we must meet with regards to managing digital content for long-term preservation, but these can be accomplished within a distributed network environment.

Implementing this strategy requires an investment in a distributed array of servers capable of storing and managing digital collections in a pre-coordinated manner. A single research institution is unlikely to have the capacity to operate multiple, geographically dispersed and securely maintained servers; MetaArchive enables research institutions to benefit from the shared network capacity made possible and financially reasonable through community-based work. The MetaArchive uses the open-source LOCKSS software, developed at Stanford University Libraries, for the network's base, and is layering data management tools on this foundation to accomplish its full preservation aims.

Replications and integrity

The distributed network enables each member's content to be preserved at multiple (currently, at least six) geographically distinct sites across two continents. These replicated copies do not merely function as back-ups to be consulted in the event of data loss, but rather are regularly compared with one another to ensure that data integrity remains consistent across all replications all the time.

Ingest and versioning

All content is ingested via HTTP, either through its access-based address (which can be open or secure) or through temporary mounting on a staging server (for collections that are not regularly available online). The ingest pathway works well with a wide variety of repository infrastructures, including DSpace, ETD-db, CONTENTdm, Fedora, and other leading solutions. For those collections maintained in an access-based address, the network servers that preserve the content regularly revisit that content to ingest any changes or additions made to it. These versions are stored alongside the original, and all versions are available to the content producer/designated community upon request. For staged collections, versioning is accomplished

through iterative re-mounting of the content for recrawl according to a revision schedule established by the content producer.

Migration

In the event that the MetaArchive Cooperative, following international practices and research, determines that a format type held in the preservation network needs to be migrated, all content of this type within the network will be migrated, and both the original copies and the migrated copies will be preserved in an ongoing manner.

Partnerships, Strategic Alliances, and Collaborations

MetaArchive is committed to extending its own community-based work through engaging in coordinated efforts with other preservation groups. Currently, the cooperative works closely with the NDIIPP program of the Library of Congress, is a founding member of the National Digital Stewardship Alliance (NDSA), and collaborates with the Networked Digital Library of Theses and Dissertations (NDLTD). The MetaArchive also is engaged actively in technical development projects with Chronopolis, Data-Pass, and the California Digital Library, as well as with several member libraries, including Penn State, Oregon State, and the University of North Texas.

Council of Prairie and Pacific University Libraries (COPPUL) Private LOCKSS Network

Contributed by Mark Jordan, Simon Fraser University

Another major PLN effort is the COPPUL PLN (http://coppullockssgroup.pbworks.com/). COPPUL is a consortium of 21 university libraries located in Manitoba, Saskatchewan, Alberta, and British Columbia who participate in resource sharing, collective purchasing, reciprocal document delivery, and other activities. As of October 2010, nine of these institutions are participating in the COPPUL Private LOCKSS Network.

One organizational challenge that the COPPUL PLN faces is that some of its members are very large by Canadian standards, while others are quite small. Large in this context includes the University of British Columbia (with a student population of over 50,000) and the University of Alberta (student population of over 35,000); the smallest is University of Winnipeg (student population of approximately 9,000). The most obvious implication of having members that differ so widely in size is ensuring that the cost of membership in the PLN remains equitable. The principal cost is annual membership in the LOCKSS Alliance (there is no membership fee for the COPPUL PLN itself), which is based on institution size and therefore mitigates this problem to a certain degree. Other costs incurred by participating in a LOCKSS PLN, however, tend to be roughly the same for all members, so smaller institutions pay more in relation to their overall budgets. These other costs include the hardware used for the LOCKSS servers, staff time required to participate in the PLN (for selection of material, integrating LOCKSS into existing or new preservation strategies, etc.), and miscellaneous costs such as travel to meetings.

Another organizational challenge particular to COPPUL is the PLN's relationship with Synergies, a "not-for-profit platform for the publication and the dissemination of research results in social sciences and humanities published in Canada." (http://www.synergiescanada.org/page/publishers) Although Synergies' use of LOCKSS is currently (as of September 2009) still in the planning stages, it is likely that Synergies journals will be preserved using some combination of the public LOCKSS network and a Synergies PLN. Any institution that belongs to the LOCKSS Alliance may be a member of multiple networks, which means that institutions will not have to pay a separate Alliance membership fee to belong to both the COPPUL and

Synergies PLNs or the public LOCKSS network. Nonetheless, questions of overlap between the networks remain outstanding and are being studied.

The main issue surrounding the sustainability of a PLN in a library consortium the size and nature of COPPUL is ensuring that enough members participate in it. The technical architecture of a PLN requires that it contains at least six nodes, since the integrity of the preserved content is verified using a voting mechanism based on this minimum. If two of the eight current members were to drop out for any reason, and were not replaced by new members, the PLN would be operating at its minimum size. In order for the COPPUL PLN to be sustainable, between a third and a quarter of COPPUL members need to participate. Given frequent cutbacks to university library budgets, this is a relatively high proportion.

The COPPUL PLN is the first PLN to automate the harvesting of OJS (Open Journal Systems) content for preservation. OJS is an open-source journal management platform developed and supported by the Public Knowledge Project (http://pkp.sfu.ca). Currently there are over 3000 journals using OJS. Many of the participants in the COPPUL PLN host OJS journals in support of scholarly publishing on their campuses or for scholarly societies, professional associations, and other academic publishers. The libraries that host these journals feel obligated to ensure that the content remains accessible if their OJS servers go offline for any reason. Preserving their journals in the PLN was an obvious strategy. The development of the LOCKSS plugin that harvests OJS content not only allows COPPUL libraries to preserve locally hosted OJS journals in their PLN, it expands the capabilities of the network. If an OJS journal is nominated for inclusion in the public network, the necessary plugin will have been created.

The COPPUL PLN, like most PLNs, preserves content that is unique to the network members and is not preserved in the public LOCKSS network. In COPPUL's case, this local content currently includes the electronic theses collection from the University of Saskatchewan, the Grande Prairie Historical Photos Collection from the University of Calgary, and the Editorial Cartoons Collection from Simon Fraser University. Determining the best way to preserve local content in the PLN offers some interesting opportunities for developing institutional preservation practices. For example, in the case of SFU's Editorial Cartoons collection, local staff refined their practice of putting the master TIFF file and various types of metadata for each cartoon into a BagIt package, which is optimized for local digitization workflow and submission to, and extraction from, the PLN (http://tools.ietf.org/html/draft-kunze-bagit-04). This technique could be applied to any institution's digital assets management workflow, however. As mentioned earlier in regards to the MetaArchive, there have been many important technology developments that have advanced the functionality and overall success of PLNs.

Alabama Digital Preservation Network (ADPNet)

Contributed by Tom Wilson, University of Alabama

The Alabama Digital Preservation Network (ADPNet, http://www.adpn.org/), a program of the Network of Alabama Academic Libraries, is a distributed digital preservation network for locally created digital content. It represents a low-cost, digital preservation solution for academic institutions, state agencies, and cultural heritage organizations in Alabama. ADPNet is another PLN, meaning that the archived content is accessible only to the members and only if it is needed to restore lost content. Any Alabama-based cultural heritage organization with publicly available digital assets whose activities and objectives are consistent with ADPNet's mission and principles may join the network.

Originally founded in 2006 by a two-year National Leadership Grant from the Institute of Museum and Library Services (IMLS), ADPNet is now self-sustaining. The current membership includes: Alabama Department of Archives and History, Auburn University, Spring Hill College, Troy University, University of

Alabama, University of Alabama at Birmingham, and University of North Alabama. For the first four years ADPNet required members to maintain an appropriately configured LOCKSS server, contribute digital content to the network and harvest digital content from other member institutions, join the LOCKSS Alliance, determine their rights to preserve content prior to submitting it to network, and hold the network and other members harmless. The costs for selecting and digitizing material, systems administration, and equipment upgrades were borne by the members individually. There was no ADPNet membership fee.

ADPNet is governed by a steering committee that represents the members, oversees the management and operation of the network, sets general policy, reviews and approves requests to expand the network's storage capacity, and reviews and approves applications for membership. The committee consists of one voting representative appointed by each member organization. The ADPNet Technical Policy Committee reviews the network's capacity and technical specifications and makes recommendations related to the network's hardware and software. The participants represent widely variant organizations in size, type, and governance. This attribute is at once a strength and a challenge. The diversity of participating organizations adds to the value and vitality of the network. At the same time, widely varying needs and resources present an opportunity to explore avenues of mutual benefit that neither inhibit the members who move forward at a rapid pace nor present barriers to members who have fewer assets to preserve. ADPNet is guided by the following operating principles:

- Mutual commitment to long-term preservation of critical cultural heritage content
- Collaboration to adopt policies and procedures that will sustain the network to the mutual benefit of its partners and content contributors
- Commitment to keeping overhead low and achieving low-cost preservation strategies
- A cooperative, robust, and decentralized peer-to-peer approach to selecting content of shared value, and mutual support of content with a particular value to individual institutions
- Application of LOCKSS software as the principal system for distributing copies of replicated content in secure, distributed locations over time
- Wide applicability to a range of institutions and digital content
- Commitment to storage and maintenance in migratable formats and data structures
- Commitment to high standards for metadata and content
- Ongoing exploration of projects to advance digital preservation

Over the past two years the ADPNet steering committee began exploring how to make the membership requirements and costs more attractive to smaller organizations in the state, particularly in light of economic conditions facing most institutions. Three key barriers existed to expanding the membership: 1) the LOCKSS Alliance fee structure, 2) the requirement to host a server, and 3) the costs of maintaining and expanding the network infrastructure.

The LOCKSS Alliance fee structure is based on the Carnegie classifications for higher education institutions. This model does not fit well for non-academic organizations (e.g., public libraries, county historical societies, etc.) and the minimum membership fee is not feasible for many of the potential ADPNet members, even if they do understand the value of digital preservation. The LOCKSS organization has worked diligently with ADPNet to address these concerns. They, too, see this issue having an effect on other Private LOCKSS Networks and have been eager to find a solution. To that end, they have agreed to allow ADPNet to freeze the current aggregate LOCKSS Alliance fees that are paid by ADPNet members to allow experimentation with amortizing those fees across the whole of the ADPNet future membership. This permits ADPNet to separate ADPNet membership from LOCKSS Alliance membership. The ADPNet steering committee is grateful for the opportunity to explore this option.

ADPNet's requirement for all members to host a server on the network automatically leaves behind many organizations that have valuable digital content, but little or no technical infrastructure to support servers and harvesting. While one of the goals of ADPNet is to increase the number of server nodes on the network, a complementary goal is to expand membership to organizations whose primary contribution is their content.

Originally, ADPNet was designed to have each member foot the cost for servers, system administration, and any necessary expansion or replacement. While this was a highly successful approach for getting the network off the ground, it limits future growth in membership of organizations with moderately substantial resources, a shrinking audience to be sure.

In response to these challenges, the ADPNet Steering Committee in consultation with current members, potential members, LOCKSS, and the NAAL Executive Committee discussed a variety of possible membership structures. They then drafted a proposal for modifying the ADPNet membership requirements by creating categories of membership that are based on the consumption of network resources rather than LOCKSS Alliance membership and local technical infrastructure. This plan includes funding for an aggregate LOCKSS fee and actual hardware costs over time. The proposal was approved by the NAAL Executive Council and the Advisory Committee in late October 2010. This next year will include a revision of the governance document, creation of a new member planning packet, and much individual consultation with potential members.

The new categories of membership are:

- Anchor
- Host
- Participant (large)
- Participant (small)

The anchor and host categories require hosting of a node on the network and are designed for members who have larger amounts of content and the necessary infrastructure to contribute in that way to ADPNet. The participant categories do not require hosting and offer preservation options that serve smaller- and medium- sized organizations with less content, but just as great of a need. Members in each category participate in the governance of ADPNet. Each category has a base annual fee that includes an allotment of space on the network with additional annual fees for incremental growth beyond the base allotment. The fee structure is designed to provide a smooth avenue of growth for members over time should they wish to move to another category. Structuring membership in this way permits organizations of any size or mission to have reasonably priced access to digital preservation in a manner that grows with their respective needs. It also adds to the sustainability of the network by expanding its reach, increasing the nodes, and stabilizing its financial business model.

By lowering the bar for participation in ADPNet even further, the long-term viability of ADPNet continues to look promising. The network provides a useful and proven option for Alabama institutions to entrust with important digital collections. The geographic distribution of participants from around the state reduces concerns regarding the reach of natural disasters. In addition, ADPNet and COPPUL are working to improve the geographical diversity of both networks by hosting one of each other's nodes.

Currently, the network preserves 138 digital collections totaling about 2.3 TB of harvested data. The current network capacity is 8 TB. Members have been focused on adding new content over the past year. ADPNet will be testing the new membership model over the next couple of years.

HathiTrust

Contributed by Jeremy York, HathiTrust

HathiTrust is a large-scale digital preservation repository that was launched by a partnership of major research libraries in 2008. There are currently more than two dozen partners (a full list can be found at (http://www.hathitrust.org). Partnership is open to research institutions worldwide. The initial focus of the partnership is on preserving and providing access to book and journal content digitized from partner collections through a number of means, including digitization by Google, the Internet Archive, and through local initiatives. The partners aim to build a comprehensive archive of published literature from around the world and develop shared strategies for managing and developing their digital and print holdings in a collaborative way. The primary community that HathiTrust serves are the members (faculty, students, and users) of its partners libraries, but the materials in HathiTrust are available to all to the extent permitted by law and contracts, providing the published record as a public good to users around the world.

Description of approach to organizational management

Organizational Status

HathiTrust is a partnership of libraries and does not have legal standing outside of those libraries (it is not a separate 501(c)3). While such standing may be helpful for some of the partners' work in the future, the partners feel strongly that their mission of ensuring the persistence and availability of the published record should be and will be most successfully fulfilled by libraries themselves, not through a separate organization.

Planning

The initial phase of HathiTrust is a five-year effort, which began in January 2008. HathiTrust will undertake a formal review of the governance and sustainability in 2011.

Finances

HathiTrust is funded in large part by the University of Michigan and Indiana University, with significant support by the Partnering Institutions and library consortia. HathiTrust does not have levels of partnership (all partners are partners with equal standing), but there are two different fee models associated with partnership. In one model, partners libraries pay for the basic infrastructure costs of the content they deposit from their digital collections. In the second model, partners may or may not contribute content, but in either case pay to support the curation of public domain volumes in HathiTrust, and in copyright volumes in HathiTrust that overlap with volumes held in their print collections. More information about these models is available at http://www.hathitrust.org/cost.

Governance

HathiTrust is governed by an executive committee and a strategic advisory board. The daily activities of HathiTrust are managed by the executive director of HathiTrust, who is appointed by the executive committee. The two bodies of governance are defined as follows:

> The executive committee manages HathiTrust's budget and finances and is responsible for final decision-making across all aspect of the partnership. It is composed of senior officers from the founding HathiTrust institutions, who will serve through the duration of the first five-year period of HathiTrust (i.e., through 2012). A listing of executive committee members, meeting minutes, and formal

charge can be found at http://www.hathitrust.org/xcom. The executive committee actively seeks input from partnering institutions via the strategic advisory board.

The strategic advisory board (SAB) advises the executive committee in areas of policy, repository development, and strategic planning. It is composed of senior professional staff from the member libraries, including four members from CIC institutions and three from the University of California. The current membership is chartered through 2012. The listing of SAB members, meeting minutes, and formal charge can be found at http://www.hathitrust.org/sab.

As part of the formal repository review in 2011, HathiTrust will convene a constitutional convention of current partners. This convention will decide future directions of HathiTrust and is the mechanism by which new partners will have a voice in determining the governance model for HathiTrust beginning in 2013. All partners who are members of HathiTrust by October 31, 2010, will participate in this convention and have a role in shaping the next phase of the partnership.

Working Groups

HathiTrust's employs a number of committees and working groups, some of which are standing and have long-term appointments, and others that may be devoted to finite tasks with a specific timeline. In general, the executive director, in coordination with the executive committee, appoints committees and working groups with an operational focus, while the SAB appoints committees and working groups with a planning or exploratory focus. SAB committees and working groups make recommendations that are reviewed by the SAB and often subsequently the executive committee, and these recommendations, when approved, may call for subsequent implementation either by existing operational groups or newly established ones.

Brief description of technical achievements

Repository

- Redundancy
  - Two complete, geographically separated, active storage sites for HathiTrust volumes functioning with load balancing and failure (a third copy of the repository is backed up on magnetic tape).
  - Redundancy of access applications including the page turner mechanism, collection-building features, and full-text search (both the full text search index, and web application).
- Size—6.6 million volumes as of mid-September with 1.3 million in the public domain. HathiTrust currently have 475 TB of usable storage at each site.
- Ingest
  - Ingest for Google and Internet Archive-digitized materials at no cost to partners. Specifications for ingest of Internet Archive-digitized volumes were developed largely by California Digital Library and the mechanism for ingest was in place in April 2010. HathiTrust has had a mechanism for ingest from Google since its inception.
  - Ability to ingest 600–700,000 volumes per month. HathiTrust has averaged more than 200,000 new volumes per month in 2009 and 2010.
- Development Environment—In December 2009, HathiTrust put out a limited release of a collaborative development environment for HathiTrust partners to work collectively on building services and applications for the repository. As of September 2010, code for existing HathiTrust applications and services has been moved to the new environment and an active development effort is in the migration

process. Current development includes an improved interface and enhanced functionality for viewing and navigation HathiTrust volumes.

Services

- Bibliographic catalog (temporary, pending release of permanent catalog being developed collaboratively with OCLC, but fully functional).
- Full text search of the entire repository (public domain and in copyright) on an ongoing basis as the repository grows.
- Collection-Builder—A features that allow users to save volumes from HathiTrust into permanent collections, which can be shared easily with others. The full text of volumes within a collection is searchable independently of the whole repository.
- Rights Management—HathiTrust manages rights information for every volume in the repository, determined through both automated and manual processes, in a rights database. The database is connected with all external-facing applications and determines access and use permissions for users accessing volumes in the repository.
- Authentication—Partner authentication is enabled via Shibboleth, increasing the services available to partner institutions.
- PDF download—Partners are able to download a full PDF for all public domain works in HathiTrust. Page at a time PDFs are available to all users, as are full PDFs of works not digitized by Google.
- Section 107 and 108 uses—The University of Michigan is making the full text of all volumes that have been scanned from its collection, both public domain and in copyright, available through a specialized interface to users who are registered at the university as having a print disability. Michigan is also providing digital access to works in its collection that are not fit for circulation because of condition or status (lost or missing). The partners are working to expand Section 107, 108, and other legal uses of materials in HathiTrust at other partner institutions.

Current initiatives

- Sustainable ingest of locally-digitized content from partner institutions. Partners have a wealth of content not digitized through large-scale processes such as Google or the Internet Archive. Variation in how volumes have been digitized and curated introduce challenges for large-scale ingest. HathiTrust is developing the policies, specifications, and processes to bring locally-digitized volumes into the collection on a large scale in a sustainable way.
- Expanding Section 107 and 108 uses.
- Open Access—HathiTrust recently introduced Creative Commons licenses in the permission agreement available for rights holders to open access to volumes. We have also begun working with some university presses to open access to volumes they have published in HathiTrust.
- Collections—HathiTrust recently charged a standing committee on collections to work specifically on issues related to building and managing the HathiTrust collection. This includes identifying opportunities for collaboration with other organizations and entities and establishing strategic directions for the partnership to pursue. The charge for the collections committee is online at http://www.hathitrust.org/wg_collections_charge.
- Developing a permanent bibliographic catalog with OCLC.
- Developing additional capability (such as bibliographic faceting) in full text search.
- Enabling computational research on HathiTrust collections through a combination of 1) distribution of datasets, 2) protocol-based access to collections, and 3) the establishment of a HathiTrust Research Center

- Quality—HathiTrust is engaged in an ongoing basis in improving the quality of volumes in HathiTrust. In addition to working closely with Google at micro and macro levels to resolve issues of quality, the partners are investigating ways to certify volumes in the repository as being fit for certain purposes (reading, printing on demand, performing computational research, etc.). This would aid librarians and users in general in making decisions about how volumes could be used (e.g., for research purposes, for collection development, and for management decisions, etc.).
- Improving communication about HathiTrust to a variety of audiences. A working group on communications was recently charged: http://www.hathitrust.org/wg_communications_charge.
- Improving usability of HathiTrust applications and services. A usability working group was also recently charged: http://www.hathitrust.org/wg_usability_charge.
- Supporting formats in the repository beyond books and journals (such as images and audio) and born-digital works.
- TRAC compliance – HathiTrust underwent an audit by the Center for Research Libraries for compliance with the Trustworthy Digital Repository Audit and Certification criteria from December 2009 through June 2010. It is are awaitng the results of the audit, which should be announced in mid-October 2010.
- Preparing for the review of the repository and constitutional convention of partners in 2011. Three partners have joined HathiTrust since it was launched in 2008 (Columbia Unviersity, New York Public Library, and Princeton University). HathiTrust expects as many as 10 new institutions to join leading up to the end October, the date by which institutions must join to participate in the constitutional convention.

DuraSpace

Contributed by Michele Kimpton and Carol Minton Morris, DuraSpace

DuraSpace (http://duraspace.org) is an independent 501(c)(3) not-for-profit organization born from a vision to help save the shared scholarly, scientific, and cultural record. It is dedicated to sustaining and improving Fedora and DSpace, two of the most dominant open source repository solutions with nearly 1000 installations of the Fedora and DSpace repository solutions worldwide. DuraSpace is also expanding its technology portfolio to respond to the web and emerging cloud opportunities to provide long-term, durable access to digital assets. The newest technology, DuraCloud, will enable digital preservation support services in the cloud. In addition, DuraSpace continues to explore new strategies for managing the data deluge by addressing the challenge of converting the overwhelming amount of data produced by scholars and scientists into useful information.

The technology portfolio inherently addresses the issue of durability of digital content. DuraSpace considers durability to be a necessary pre-requisite to the process of digital preservation. It is especially interested in providing technologies and services that ensure that digital content is accessible over the long term. Thus, durability translates to long-term or perpetual access to digital content. The organization has adopted the organizational byline, “open technologies for durable digital content.”

Description of approach to organizational management

Organizational Status

The executive team of DuraSpace is comprised of the Chief Executive Officer (Sandy Payette), the Chief Business Officer (Michele Kimpton), and the Chief Technology Officer (Brad McLean). Together the executive team sets the strategy and provides management and oversight for DuraSpace organization, bringing forth the assets and experiences from both Fedora Commons and the DSpace Foundation. The

DuraSpace organization employs technical leaders and community outreach specialists to managed its portfolio of open source projects and to support the very active communities that surround them. Software developers are on staff to work on new technology initiatives and grant-related deliverables.

The DuraSpace organization works with a board of directors to develop strategies to capitalize on the strengths inherent in its technology portfolio and to utilize viable business models for the long-term sustainability of the organization. A key focus of the work with the Board in 2010 has been on refining DuraSpace's strategic positioning, implementing business models, and securing revenue sources to fulfill its mission and achieve its financial goals for the five-year forecast and beyond.

Finances

DuraSpace's funding strategy has three prongs: (1) endowment and grants, (2) income from sponsorships and programs, and (3) revenue from services. The endowment fund originated with a significant grant from the Moore Foundation. We continue to secure new grants that are targeted at new community programs and innovations in DuraSpace's technologies. The annual Sponsorship Program brings in support from the wider DuraSpace community. Other programs also help support the organization such as the Registered Service Provider Program. The newest prong of the funding strategy is revenue generated by services. The first service that will be introduced is the DuraCloud hosted service, providing pay-for-use, cloud-based services supporting digital preservation and access. DuraCloud has been released in open source, and the hosted service is scheduled to launch in Q1 2011. In the future, DuraSpace will offer other types of services that are of value to its communities, including education and training.

Community

The DuraSpace organizational context provides a vast network of community collaborators via its existing open source projects. Since end user involvement is essential to its work, DuraSpace leverages partnerships that enable it to collaborate with domain scientists, researchers, librarians, and data specialists to address the challenges of data management, data curation, and digital preservation.

Communities are essential to the creation of shared and enduring solutions. To address the challenges inherent in preserving and providing access to digital information, the power of community can be harnessed to protect digital heritage, promote open access, enable scholarly communication, manage scientific data, and more. Community groups include user groups, committers groups, solution communities, the DSpace Global Outreach Group, and others that begin with a bottom-up approach to community organization that enables grassroots efforts to flourish.

DuraSpace supports the community with a variety of web-based tools and services (publications, webinars, community mailing lists/forums, online user registries, user group meetings and more) to keep users up to date with the latest news and ideas from around the globe.

Brief description of technical achievements

Fedora and DSpace have their origins in research programs based originally at Cornell University, University of Virginia (Fedora), and at MIT with Hewlett Packard Labs (DSpace), with funding by NSF and others. From these roots, both have emerged as mature, open source projects with established communities and well-honed software development processes. Collectively, both are deployed worldwide, with a total of over 900 installations supporting digital preservation, digital libraries, digital archives, virtual research environments, and open access platforms.

DSpace

DSpace (http://www.dspace.org/) is an out-of-the-box, open source repository application that is end-user oriented, providing user interfaces with data submission workflows and a range of features supporting both preservation and web-based access to digital content. There are over 750 digital repositories using DSpace software. Globally, it is the most widely used open source repository software for institutional repositories and open access repositories. DSpace is typically used by organizations, especially university libraries, as a way to provide access to research output, scholarly publications, and digital collections.

The DSpace application has many features and tools for managing digital content and enabling digital preservation. Organizations can easily make their digital collections available on the web using DSpace's customizable end user interfaces, along with many community-developed features and utilities. DSpace 1.6 was released in March 2010. The DSpace community has a long history of community involvement, with an international development team of open source committers and hundreds of contributors.

Key Features of DSpace:

- Out of the box: DSpace is easy to install and get up and running quickly.
- Built-in workflows: Originally designed for libraries, the embedded DSpace data model and workflows are familiar to librarians and archivists.
- Built-in search engine: DSpace supports full-text searching for end users.
- File types: DSpace supports all major file formats.
- Security: The platform comes with an authorization stack, or organizations may use an existing LDAP or similar protocols to link their internal systems.
- Permissions: DSpace has access control as granular as item level, or you can set global permissions based on communities and collections.
- Many community-developed tools and add-ons.
- Standards compliant: DSpace complies with OAI-PMH metadata harvesting and SWORD ingest.

Fedora

Fedora (http://www.fedora-commons.org/) is a digital repository system used to manage diverse collections of scholarly, scientific, historic, and cultural materials. From its origins at Cornell University, Fedora has evolved into mature, open source software that provides a robust, modular system for the management and dissemination of all types of digital content. Its flexibility and service-orientation has enabled it to integrate gracefully with many types of data management environments including enterprise, web-based, and grid-based systems. Its scalability is well documented, with test results for 10 million to 100 million objects. There are over 175 institutions registered with Fedora installations in universities, research institutions, research libraries, national libraries, non-profit organizations, and government agencies. Fedora 3.4 was released in June 2010. The community-based development team is international, and currently consists of 12 open source committers and many contributors.

Fedora provides a flexible model for digital preservation and archiving. Fedora digital objects are self-describing, meaning that all essential characteristics of the object are packaged within. All Fedora data management and access functions are exposed via well-defined web APIs. Also, the repository is integrated with the Mulgara triplestore (http://mulgara.org) and has query capabilities using the semantic web SPARQL query language.

Key Features of Fedora:

- Store all types of content and its metadata
  - Digital content of any type
  - Metadata any format
  - RDF relationships
- Scale to millions of objects
- Access and management via web APIs
- Ability to integrate with multiple, customer-driven front ends
- Disaster recovery features
- Content models (define "types" of digital objects)
- Storage plug-in architecture
- Authentication policy enforcement
- Web-based administrator client
- OAI-PMH provider service
- Full-text search

DuraCloud

The new DuraCloud technology (http://duracloud.org/) is both a hosted service and open technology developed by DuraSpace that makes it easy for organizations and end users to use cloud services. DuraCloud leverages existing cloud infrastructure to enable durability and access to digital content.

DuraCloud will be deployed as a service hosted by the DuraSpace organization using a cloud server environment and is integrated with multiple cloud storage providers, including Amazon AWS and Rackspace.

DuraCloud has been deployed already as open source software in June 2010. The decision to open source the code base was made to encourage community involvement in the development of the software, as well as to allow the creation of new services which can be integrated with the DuraCloud system.

The following features are significant for supporting digital preservation and data curation:

1. **Replication Service**: Transparently push content to multiple, third-party storage providers so all users can take advantage of low-cost, internet-based storage. A major benefit of DuraCloud is risk mitigation by mediating to multiple clouds to support multiple copies of content and to diversify over different cloud providers.
2. **Fixity Service**: In order to ensure that data in the cloud remains viable, this service allows users to check the bit integrity of content stored within DuraCloud and includes many configuration options to fit various usage needs.
3. **Bulk Image Conversion Service**: This service handles the conversion of large numbers of image files from one format into other. The server process management is handled by utilizing Hadoop, which allows the conversion to run over multiple servers of varying sizes, thereby increasing the overall throughput. Users with a large number of images stored in DuraCloud will find this service particularly useful.
4. **Sync Tool**: The Sync Tool allows users to easily move content from the local system (such as a DSpace or Fedora repository) into DuraCloud. This new feature allows the Sync Tool to exit when all files are synchronized, rather than continually monitoring the system for changes.

5. **Multi-Delete Feature**: To make managing large and complex data sets easier, DuraCloud now provides a simple way to delete groups of content items and spaces in one step.
6. **Media Access Services**: DuraCloud provides specialized access services to view JPEG 2000 images and stream media such as video and audio.

Current Initiatives

1. Work with board of directors, Gold Sponsors, and key community organizations to develop synergies and alliances in support of shared solutions and new funding models for digital preservation.
2. First Annual Gold Sponsor Strategic Forum to work with key leaders and supporters on strategic issues and future directions for DuraSpace.
3. Launch the DuraCloud public release in Q1 2011.
4. Provide technical leadership in the digital repositories domain by enabling new system integrations with DSpace, Fedora, and other open source community projects.
5. Market study of research organizations who are using DuraSpace technologies to discover how we can better serve the researcher community.
6. Grow the Registered Service Provider program to improve collaboration between DuraSpace and its network of service providers.
7. Broaden and expand community programs that provide new resources, services, and educational opportunities that enable the Fedora and DSpace communities.

Data-PASS: The Data Preservation Alliance for the Social Sciences

Contributed by Micah Altman, Institute for Quantitative Social Science, Henry A. Murray Archive, Harvard University

The Data Preservation Alliance for the Social Sciences (Data-PASS) (http://www.icpsr.umich.edu/icpsrweb/DATAPASS/) is a broad-based voluntary partnership of data archives dedicated to acquiring, cataloging, and preserving social science data, and to developing and advocating best practices in digital preservation. Collectively, the founding partners have over 200 years of combined experience in social science data archiving. These partners include the Inter-university Consortium for Political and Social Research (based at the University of Michigan), the Roper Center for Public Opinion Research (based at the University of Connecticut), The Howard W. Odum Institute for Research in Social Science (based at the University of North Carolina at Chapel Hill), the electronic records custodial division of the National Archives and Records Administration (NARA); and The Henry A. Murray Research Archive (based at Harvard University).

An award from the US Library of Congress' National Digital Information Infrastructure and Preservation Program (NDIIPP) catalyzed the Data-PASS partnership. The partnership is a founding member of the new National Digital Stewardship Alliance, a collaborative effort among government agencies, educational institutions, non-profit organizations, and businesses to preserve a distributed national digital collection for the benefit of citizens now and in the future.

As of 2010, the partnership has identified thousands of at-risk research studies and acquired over 1300 studies. Data-PASS also has established a shared online catalog for the tens of thousands of studies or series that comprise each partner's entire data holdings.

The preservation of quantitative data has a more extensive history and more well-established practices than in most other disciplines. Social science continues to rely heavily on data in its traditional forms, such as opinion polls, voting records, surveys, and government statistics and indices. On the other hand, although most large data sets are in public archives, most data produced by and used in social science research is

neither publicly available nor preserved by an archival organization. Digital content is evolving into more forms than can be preserved readily. Changes in technology and society are greatly affecting the types and quantities of potential data available for social-scientific analysis. Any data describing human activity may be a subject of social science research. Taken as a whole, the evidence base of social science is shifting, and consequently, approaches to curating this evidence, or data, is shifting as well.

Data-PASS is currently engaged in two technical initiatives. One major initiative by the partners is the development of the SAFE-Archive system. This open source system will provide verified, policy-based, distributed replication for the Data-PASS network. The system will allow any library, museum, or archive to easily replicate their content and to audit the replicas of that content using an existing public or private LOCKSS network.

The archival community has largely recognized that a geographically and organizationally distributed approach is necessary to minimize long-term risks to digital materials. The innovation of the SAFE-Archive system is that it automates policy—the behavior of the syndicated storage system is audited automatically by reference to an archival policy. The Data-PASS members describe these policies formally using a metadata schema. The formal policies include systematically describing the commitment of resources each of the archives has made to preserve the contents of the other partners, the auditing commitments each has made to its depositors, and the legal policies supporting access to the data by other partners in the case of institutional failure. Furthermore, commitments may be asymmetric, to facilitate partnerships among institutions of different sizes. The system acts on this metadata by auditing the actual state of the replication network and reporting any deviations from policy. These auditing reports are schematized so they can be machine manipulated and explicitly organized to link to and support TRAC criteria. The system will help smaller library museums and archives to establish compliance with an important part of emerging “trusted repository” standards. By schematizing the policies, “trusted repository” requirements will be measurable, actionable, and auditable.

A second Data-PASS initiative is the promotion of citation standards for research data. Accurate citation of data will promote more and better science. It will make data easier to find, to replicate, and to manage for the long term. Moreover, it will make it much easier to trace the influence of data on social science. Data-PASS uses a simple baseline standard: title, author, data, and a persistent identifier (of any widely recognized type, such as URN’s, handles, DOI’s, etc.). Data-PASS also recommends the addition of fixity information, such as a checksum or Universal Numeric Fingerprint [CITE], which verifies that data used later matches data originally cited.

The Data-PASS partnership has been conducting outreach to professional societies, encouraging the use of a lightweight data citation format. Much of the gains from the data citation can be realized by including these simple elements in a single consistent place in publications. To encourage data citation by journals, the partners offer free or low-cost permanent archiving for publication related data. Further, the Dataverse Network System (King 2007), which provides Data-PASS’ unified catalog, metadata, and data preservation and dissemination services, can be used to create free digital archives for journals that wish to have a dedicated replication archive under their curatorial control using their branding.

Organizational Structure and Ongoing Collaborations

The partnership is voluntary and is structured by a formal memorandum of understanding—this is available on the partnership website. Data-PASS partners make a commitment to ratify the membership agreement, which gives the partnership the licenses necessary to expose a member’s content through the catalog, and to replicate it for preservation. Each Data-PASS member retains full ownership of its content and licenses Data-PASS to harvest it for inclusion in the catalog and preservation system. Each partner also attends virtual meetings of the steering and operations committees.

There are no membership fees. Partners contribute in-kind personnel time and computing resources to keep overhead low. Partners also collaborate on grant proposals or other fundraising efforts to cover the costs of developing new tools and conducting research into archival practice.

The University of California Curation Center, California Digital Library

Contributed by Patricia Cruse, Stephen Abrams, and Perry Willet, University of California Curation Center, California Digital Library

Digital information is vital to the University of California's research, teaching, and learning mission. The digital environment has fundamentally transformed the way in which information is produced and disseminated within the university, blurring the lines between knowledge creation and formal publication; changing the way users find, access, and use information; and creating new demands for effective curation of digital content in a wide variety of formats. Within the UC system the newly established UC Curation Center (UC3), one of five programmatic units of the California Digital Library (CDL), has a broad mandate to provide innovative solutions that ensure the long-term usability of the university's digital assets (http://www.cdlib.org/uc3).

UC Curation Center, a Creative Partnership

As a central, system-wide service provider to the ten UC campuses, UC3 is routinely asked to assume custodial stewardship for digital content in ever increasing number, size, and diversity of type. Furthermore, this content is often used and repurposed in novel contexts far removed from the intention of its original creators. Thus, the programmatic imperative of UC3 is to provide a curation environment that is comprehensive in scope, yet flexible with regard to local policies and practices, responsive to requirements by funding agencies for data management and open access, and cognizant of the inevitability of disruptive changes in technology and user expectation. By bringing together the experience, expertise, and resources of the CDL, the ten UC campuses, and the broader international curation community, the UC Curation Center fosters collaborative analysis, projects, and solutions to ensure the long-term viability and usability of curated digital content. Harnessing the collective energy and innovation of its partners, UC3 provides solutions that are out of the reach of any individual partner.

UC3 works through this partnership to:

- build a community of shared concern and practice
- create channels to pool and distribute diverse experience, expertise, and resources
- provide robust, innovative, and cost-effective solutions to counteract inevitable disruptive change

This collaborative model enables the UC libraries and the campus communities to act as cost-effective guardians over UC's scholarly digital assets without having to invest individually in the requisite deep technical expertise and infrastructure. As founding partners of the Center, the UCLA Library, the UC San Diego Libraries, and the UC Merced Library have helped to shape the UC3 organization. UC3 promotes and cultivates partnerships across the University and beyond, leveraging its expertise and technology to meet systemwide needs. The UC3 is building a growing community of faculty members, researchers, librarians, archivists, curators, IT professionals, and administrators by bringing together everyone with a stake in the ongoing viability of digital information. The Center is built on innovation, creativity, trust, and excellence, and is the hub of digital preservation and curation activities for the University of California.

New Technical Approach, Micro-Services

To respond more efficiently and flexibly to the changing information landscape, UC3 has developed a new technical infrastructure built on the concept of "micro-services." Micro-services are an approach to digital curation based on devolving curation function into a granular set of independent, but highly interoperable, services that embody curation values and strategies (http://www.cdlib.org/uc3/curation). Since each of the services is small and self-contained, they are collectively easier to develop, deploy, maintain, and enhance. Equally as important, they are more easily replaced when they have outlived their usefulness. Although the individual services are narrowly scoped, the complex global function needed for effective curation is nevertheless an emergent property of the strategic combination of individual services.

Micro-services can be deployed in the environments in which it makes most sense, both technically and administratively. While UC3 will continue to use micro-services as the basis for Merritt (http://merritt.cdlib.org/), its centrally-managed curation repository service, micro-services-based systems can also be deployed and operated in local campus environments. With micro-services it is no longer necessary that digital content must be transferred to a common repository in order to receive appropriate curation care.

The UC Curation Center's Rich Service Offering

UC3's organizational model combined with its new technical approach allows us to offer a rich range of services—from guidance and best practices to technical infrastructure and hosted solutions—that are designed collectively to respond efficiently and effectively to the challenges posed by the information landscape. The following is a sample of UC3's rich service offerings.

Consultation Services

Given the complex landscape of digital curation and preservation, the UC3 has put in place a consultation framework to help its community of users find the most appropriate solutions on issues involving creating, managing, and preserving digital objects to the challenges they face. Specifically, UC3 provides a "Data Management Service" that provides expert advice on managing data throughout the research life cycle to ensure usability, preservation, and access (http://www.cdlib.org/services/uc3/datamanagement/). This is a key service for UC faculty and researchers as federal agencies and other funders are now requiring that grant awardees include a data management and sustainability plan with their grant proposals.

Hosted Services

UC3 hosts and maintains a range of services for use by UC community and beyond including:

a. **Merritt** (http://merritt.cdlib.org), a new, cost-effective repository service that lets users manage, archive, and share their valuable digital content. Built using UC3's micro-services, Merritt provides significant features for digital content management: permanent storage, access via persistent URLs, tools for long-term curatorial management, and an easy-to-use interface for deposit, update, discovery, and retrieval. Merritt is open to all disciplines, has no restrictions regarding acceptable formats, and has its full functionality available via intuitive, web-based user interfaces as well as programmatic APIs.
b. **EZID** (http://www.n2t.net/ezid) (easy-eye-dee), a service that makes it simple for digital object producers (researchers and others) to obtain and manage long-term identifiers for their digital content. The service can create and resolve identifiers on behalf of the user and also allow the user to enter and maintain metadata information about the identifier and the content that it identifies. EZID helps researchers take control of the management and distribution of their research, share and get credit

for that research, and reap the reputational benefits through its collection and documentation. EZID makes objects easier to access, use, and reuse. As a result, it also makes it easier to build on previous work, conduct new research, and avoid duplicating previous efforts. The service is available via an intuitive web-based user interface as well as a programmatic API.

c. **Web Archiving Service (WAS)** (http://was.cdlib.org), a system that provides tools to collect, manage, preserve, and publish web content. The web has revolutionized our access to information, but web-based publications are inherently fragile over time, and ready access to these resources cannot be taken for granted. The Web Archiving Service enables librarians and scholars to meet that challenge.

Community Initiatives

The landscape of digital preservation and curation is large and complex, and it demands a range of solutions and partners. UC3 is deeply engaged with a number of partners to provide the highest quality and most cost-effective solutions for digital curation to the UC community.

a. **Chronopolis** (http://chronopolis.sdsc.edu/) is a national center for the management, long-term preservation, and promulgation of digital assets, based at the University of California San Diego, where it is co-managed by the UCSD Libraries and the San Diego Supercomputer Center. Funded initially by the Library of Congress' National Digital Information Infrastructure & Preservation Program (NDIIPP), Chronopolis provides preservation support for a wide range of collections around the United States. A number of the tools used within Chronopolis have been developed in conjunction with UC3 staff at CDL.

b. **DataONE** (http://www.dataone.org) is a project to ensure preservation and access to multi-scale, multi-discipline, and multi-national science data, funded by the US National Science Foundation. Researchers at the University of California have partnered with dozens of other universities and agencies to create DataONE (Data Observation Network for Earth), a global data access and preservation network for Earth and environmental science that will support breakthroughs in environmental research. DataONE is one of two $20 million awards made this year as part of the NSF's DataNet program. On behalf of UC libraries, UC3 is an active member of DataONE, which is developing the cyberinfrastructure and organization to support the preservation and access to climate change data. By working collaboratively with UC Davis, the UC Santa Barbara Library and National Center for Ecological Analysis and Synthesis (NCEAS), and a range of other partners on this large scale international initiative, UC3 works to ensure that the needs of UC's research community are investigated by the project partners and that services developed through that investigation are appropriate and available to the UC community.

The UC Curation Center offers a range of innovative and sustainable curation services to its stakeholder communities. The technical underpinning of these services is based on the micro-services approach to curation infrastructure, which emphasizes the decomposition of curation function into a growing set of individual micro-services that can be used to compose a variety of useful systems. The granular decomposition of function facilitates the independent enhancement and replacement of component parts without effecting global curation service availability and behavior.

The organizational structure of the Center, which draws on the combined expertise, experience, and resources of the California Digital Library and the ten UC campuses, facilitates information interchange, pools and distributes innovative experimentation and approaches, and enables a range of both centrally supported and locally provided solutions that can be tailored to the particular needs of the Center's

customers. The combination of a robust technical infrastructure and a flexible organizational structure enables the UC Curation Center to respond and evolve quickly, efficiently, and effectively to the changing curation needs of the University of California.

4

Concluding Recommendations

What Can Your Library Do?

As posited at the beginning of this report, research libraries are at a critical juncture with regard to their present and future roles. They must carefully consider whether they intend to be lifecycle managers of the new forms of digital information objects created by members of their parent institutions. The growth rate for production of such intellectual objects is increasing at an alarmingly fast pace, and if research libraries do not actively take on this responsibility, other entities, both on campus and beyond, will step in to do so. We already see this happening.

The following recommendations highlight the institutional sphere of engagement but should not be read as preferencing "go it alone" strategies. Wherever possible, collaborative or community-based approaches to digital curation are likely to be more effective and sustainable. However, effective use of collaborative strategies and community-managed resources requires local investment and capacity building. The following recommendations reflect this reality. How should individual libraries conceptualize themselves in this quickly changing terrain? And what should individual institutions do in order to position themselves for success in their campus settings as well as within the broader digital library field? We recommend the following:

- **Stop waiting and start proactive engagement locally**
  This should be done through relationship-building with the Office of Sponsored Programs, the Provost, the Vice President for Research, the Deans and department chairs, through other grant connections, data centers, and individual professors and researchers in your institution. Faculty partnerships are a significant driver in digital curation initiatives as are policy partnerships with campus offices.

- **Stake a claim in the production cycle**
  Consider new services in creating multimedia and other digital assets by hosting e-publishing services (e-journals, e-books, e-conference proceedings, etc.), and by hosting and curating digital archives, datasets, digital art objects and websites, etc. If staffing these initial services is not possible currently, then look into utilizing graduate students skilled in these areas and a fee-based model for recovering costs. The important step is to bring these services in-house and seize the moments of opportunity as they arise.

- **Start retraining and repurposing staff**
  Very few research libraries should have more than half of their infrastructure devoted to physical collections at this point in time. The library needs to think of digital curation as a core function of the library and to invest financial and other resources into it accordingly. Seek in-depth, long-term training programs for your interested staff. Bring experts to your library. Maintain the daily conversation that your library is and will be engaging in digital curation services.

- **Be a doer, not a broker, wherever possible**
  Think about the implications of outsourcing before making decisions that might hurt in the long run. We must ask ourselves, what are the core functions of a research library? Then, we must avoid outsourcing

our reason for being—the functions for which we exist and in which must excel. Do not surrender ownership of materials; do not surrender infrastructure. There is no need to hand over our future to external groups; research libraries have adequately demonstrated that it is possible to collaborate in managing digital content and maintaining staff and technology infrastructures in an economically viable way.

- **Consider digital curation collaborations**
  Which research institutions have you collaborated with before? Ask yourself if they have the potential to be partners in erecting digital curation services and technologies? Did they deliver in past projects? Do you have a positive rapport with them? Do you trust them as partners? Together, pick a project or a meaningful aspect of technology infrastructure and begin building a long-term relationship.

- **Actualize collaborative engagement**
  Following from this examination, work collaboratively and steadily with other selected institutions over time to build a sustainable cyberinfrastructure. Early experiences and trends indicate that multi-institutional collaborative approaches provide a successful organizational context that is necessary to meet this large and ever-shifting challenge.

Research libraries have much to gain if they behave as active players in the development and management of research cyberinfrastructure across the arts, humanities, social sciences, sciences, and engineering fields. If they embrace their emergent roles as anthropologists of research environments, co-producers and broadcasters of digital content, and builders of systems for research collaboration, communication, and scholarly object management, then research libraries and their employees will successfully transform themselves into highly regarded service entities and partners in the global digital research community of the early twenty-first century. In order to pursue this direction successfully, research libraries will need to shift their financial and human resources toward it. In so doing, they will be choosing a path that requires them to rise to the occasion and invest in this developing path of cyberinfrastructure and services. If the shifts in investment are half-hearted, then research libraries risk losing their potential to become a sought-after service entity in the digital research and scholarship domain. However, if they choose to embrace this path, their growth and indispensability as partners in digital research and scholarship may be limitless, as the digital realm itself grows in seemingly endless ways.

Once the choice to walk the digital curation pathway has been made, research library leaders and their staff should acknowledge that we now have many tried and tested cyberinfrastructure models that we operate ourselves. These models and technologies—inter-institutional and community-driven—have experienced early successes, and there is no reason that they cannot continue to do so. Research library leaders are often motivated by what seem like easy solutions brought to them by many types of vendors. However, our history is littered with examples of libraries jumping to incorporate new commercial software and services, only to see the institutions they serve become dependent on their closed technological approaches. From that dependent space we have watched companies experience early demise because they become financially unsustainable, and on the flip side we have watched successful companies raise their prices and change their policies in ways that have been detrimental to our community. The commercial model is not always the correct or most appropriate model, and through this report we strongly urge research libraries to consider longer-term, more stable and sustainable models that they invest in and own when it comes to digital curation for preservation.

We are at a critical juncture in the life of digital curation for preservation services, where research libraries in the aggregate will determine if this domain goes commercial or if it stays predominately inter-institutional and community-driven between research institutions. We must find meaningful roles for the commercial entities to play, but they must play within the frameworks of research libraries' community-driven approaches if we are to control costs, maintain and evolve technologies properly, and perhaps most importantly, sustain the digital data, information, and knowledge objects that have critical intellectual value.

Appendix A

Disciplinary Considerations for Digital Curation: The Sciences

Scientific and Engineering Cyberinfrastructure and Data Curation

Modern e-science is one phenomenon where research libraries have an overwhelming abundance of opportunity to engage in and serve the research process. The context for this is well described by Tony Hey, Stewart Tansley, and Kristin Tolle in their 2009 edited volume, *The Fourth Paradigm: Data-Intensive Scientific Discovery*. They put forth that the major characteristics of the fourth paradigm of scientific discovery begins with data. Scientists mine vast amounts of accumulated data collected by sensors and other recording devices and analyze it. The analyzing is done through robust computational means, often invoking the use of high-performance computing platforms to process the data and search for correlations and patterns. This digital data can take the form of numeric tables, still images, moving images, and other digital textual, numeric, and graphic representations. For research libraries, this data looks like just another set of digital files with informational content held in various file formats. But of course, it is not that simple. Robust metadata is required to document the scientific process, as well as the technologies used in producing the data. Specialized software tools may have been developed to generate, analyze, and visualize the data. These tools may need to be managed and preserved, as well. Additionally, the data is large, increasing through terabytes to petabytes, exabytes, and beyond. Despite the complexities, e-science data is another digital resource that research libraries can manage for researchers. They can facilitate its sharing, use, re-use, and preservation as researchers pursue new scientific understandings and discoveries.

Some of the potential roles for information professionals in managing and preserving content in cyberinfrastructure environments, including the sciences, engineering, and quantitative social sciences, can be summarized as follows:

1. Creation of research data and information objects
    - Data modeling, ontology and taxonomy development and application
    - Initial capture and management of data coming from instrumentation
    - Capture of content coming from research team's web-based communication tools
2. Curation and management of research data and information objects
    - Collection / ingest, description, provenance-tracking, access and reuse, integration, maintenance, and digital preservation of data and information
    - Provision of computational resources and data storage
3. Collaboration in virtual communities
    - Manage collaborative web spaces to connect researchers in community who participate in a certain area of research or research project with one another (i.e., package and deliver relevant content to the community, devise web-based social networks, assist with linking people together, and find information)

This shift of roles for information professionals, including librarians, brings them into the data and information lifecycle earlier; their work focuses on adding value in the knowledge generation process, not

just managing the information product that comes at the end of the cycle. Anna Gold has synthesized nicely some of the new services and activities:

> "Linking data in rich and robust ways to support data reuse and integration will require understanding and documentation of the data's provenance, the development of ontologies, expert annotation, and analysis. Further downstream, services enabled by these activities will include visualization, simulation, data mining and modeling, and other forms of knowledge representation and extraction." [1]

Curating these communications and information products for long-term access purposes is, in essence, applying the digital curation for preservation role that research libraries increasingly are pioneering.[2]

Creation and Curation of Research Data and Information Objects

The term data curation helps to illuminate the roles research libraries are beginning to play in the preservation of digital scientific, social science, and engineering data. Several reports have started to document the formation of this data lifecycle management role.[3]

Issues and Challenges

There are many noteworthy data curation challenges that research libraries and information professionals face and are addressing as part of their data management role. One of the main policy challenges is managing cost. The costs associated with ensuring data's accessibility for a very long time (and certainly longer than the research grant monies last) are many and high. Hence, a key component in managing cost is the appraisal of data to determine its long-term research value—locally, regionally, nationally, and internationally. However, the "devil of appraisal" [4] is in the details. Data appraisal as a manual review process is too time consuming and expensive, and automated approaches to such decision-making are in their infancy, if even feasible. The practice of data appraisal is swiftly becoming a necessity in the "data deluge" environment, and research institutions need information professionals with a toolkit of techniques to help them mitigate curation costs and operational issues. Research librarians and archivists have much to contribute, but it will take time and effort within the research library community to build the requisite body of knowledge and professional practice for success as research data curators.

Another challenge in data set management has to do with the varying bodies of expertise that need to be brought to bear on the data curation problem. Beyond information technology and library/archival science knowledge, information professionals may need a deep disciplinary knowledge to apply when appraising and curating research data, as well as when assisting in the creation of data by designing ontologies, taxonomies, and creating other forms of metadata. Bringing multiple knowledge bases together for a data curation effort means involving multiple persons and/or layers of training (e.g., a PhD in a social science field and a MLIS or other data management degree), which can equate to high expense—though not as high as the cost of either attempting to curate all data sets created within an institution *or* curating only a selection of this research output, and one that is not carefully chosen. If the research library wishes to perform this function for its parent institution, it first must justify the cultivation of such expertise in the library as a practical investment that the campus may leverage across departments.

Institutional policies regarding data appraisal (determining which data is significant enough to apply institutional resources to curating it) also lead to institutional obligations and the responsibility to maintain data, and once again, building the organizational capacity to preserve vast amounts of research data. These policy issues largely focus on resource use decisions, such as locating and dedicating funding to the effort. They also, however, include contractual issues, such as understanding obligations to agencies funding the research from which the data came. Ethical issues abound when deciding when data is curated for the long-

term, when it receives "triage," when it receives no curatorial management and basic storage only, and when it is not maintained at all beyond the life of the project.

While storage seems a basic, even mundane issue, it is actually one of the most vexing policy challenges in e-research due to the size and provisioning of data storage services. Capacity will run into the terabytes, petabytes, and eventually, exabytes. Who will pay for such huge levels of storage, create a support organization around it, and parcel it out to researchers? How can we build the organizational capacity that our campuses need in order to provide managed services for data curation? As identified by Cliff Lynch, "data storage is a shared resource involving the campus, the researcher, the funding agency, and possibly industry and other outside partners." [5] Lynch argues for incentivizing stakeholders to contribute to the maintenance of data storage services as a shared resource.[6] Business modeling, formalized trust relationships, and functional roles all need to be developed and clarified in the complex data curation environment.

Many vexing digital curation issues lie between the steps of deciding what to keep and how and where to store it. Many simple functional areas in science data curation, as well as the broader concept of digital curation—encompassing all forms of digital objects holding intellectual or academic value—need attention. Activities such as dataset ingest, description, provenance-tracking, access and reuse, and long-term stewardship all present their challenges, and many articles and reports document the professional process of analyzing them.[7] Format characterization tools like JHOVE and format registries like PRONOM and the upcoming Unified Digital Formats Registry (UDFR) are needed to affirm what, exactly, we are ingesting. It is also crucial to be able to track the provenance of data. **Data provenance** refers to the ability to trace and verify the creation of data, how it has been used or moved among different databases, as well as altered throughout its lifecycle. Data users can easily alter web-available data, making it difficult to know the data's origins and original character. Provenance is a significant issue in the process of ensuring that data is authentic and reliable to its users in the validation of scientific experiments, their findings, and conclusions. Data description is another area of much work, although projects like the University of Illinois and Purdue University's Data Curation Profiles project have advanced the field's ability to document provenance, as well as understand the content and context of data.[8] The issues of data management policy, appraisal, cost management, authenticity and reliability, provenance tracking, and storage are areas where research libraries, typically in partnership with other academic units such as information technology and the domain researchers and specialists, can play significant roles and positively impact the creation, conveyance, and preservation of unique research data in digital form.

In Response: Data Curation Programs

There are many substantive data curation initiatives ongoing in research libraries. For example, the recent e-science survey work performed by the ARL E-Science Committee, and the subsequent ARL report by Soehner, Steeves, and Ward, "E-Science and Data Support Services: A Study of ARL Member Institutions," document well the growth of library-based science data curation programs.[9] Among the disciplinary data being reviewed and/or managed are collections serving the biosciences/bioinformatics, atmospheric, environmental, and geosciences data, agricultural data, astrophysical data, and neuroscience data, among others.

Many of the research libraries involved, such as Washington, Utah, Oregon, Purdue, UIUC, MIT, and Georgia Tech, are spending considerable time and energy conducting studies of researcher data practices on their respective campuses, as well as assessing researchers' needs with regard to creating, maintaining, sharing, using, and sustaining their data. Others have created centers or institutes dedicated to researching issues of digital and data curation, fostering campus collaborations in data curation, and offering educational opportunities and awareness-building events. Purdue, for instance, has created its

Distributed Digital Curation Center, Johns Hopkins has established its Institute of Data Intensive Engineering and Science, Washington has started its eScience Institute , and Illinois has developed its Center for Informatics Research in Science and Scholarship.[10] All involve the university libraries at varying levels. These developments demonstrate significant commitment to science data-related issues and should give birth to new understanding and solutions in the broader digital curation realm, as well as the more specific domain of e-science data curation.

Some libraries have embarked on piloted services where they have involved subject librarians, who spend time discussing data needs and issues with their faculty and serve as a bridge to early data curation services (i.e., storage, retrieval/access, use and reuse, sharing with colleagues, etc.). Universities such as Cornell and MIT have utilized this approach, where subject librarians take on data management roles, largely self-taught, and their technology units design and erect incipient infrastructures that support data services using Fedora, DSpace, and other software utilities and applications to manage data.[11] Several of these initiatives maintain cross-unit collaborations on their campuses. Perhaps the best example of this approach is found at the University of California, San Diego (UCSD), where the university libraries are collaborating with the San Diego Supercomputer Center to develop and grow their campus cyberinfrastructure for e-science purposes. In particular, the two UCSD units have come together to form Chronopolis, a digital preservation service focusing mainly on science data needs, which is made available to UCSD as well as to other research institutions. Current client institutions include North Carolina State University, University of Michigan, California Digital Library, and the Scripps Institution of Oceanography at UCSD. Research, education, assessments, new infrastructures and services, and opportunities for library, IT, and faculty collaborations are very plentiful, as research libraries move forward in offering to partner up and manage this new digital information resource of import to data-driven research disciplines.

Research libraries also have been front and center in the DataNet Program, sponsored by the National Science Foundation's Office of Cyberinfrastructure. These $20 million, five-year grants promise to yield new understandings, approaches, and eventual infrastructures for science data curation. To date, two very large and collaborative projects have been funded: the Johns Hopkins University's Data Conservancy and the University of New Mexico's DataONE initiative.[12] The former builds upon Johns Hopkins' work with astrophysics data and looks to build infrastructures for all manner of sciences and engineering fields, while the latter is an "observation network for Earth," encompassing "biological data available from the genome to the ecosystem; and …environmental data available from atmospheric, ecological, hydrological, and oceanographic sources." Future projects and initiatives of this kind may be forthcoming with NSF support.

Research Library Collaboration in Virtual Communities

In science and engineering fields, many collaborative endeavors cross institutional and physical boundaries in ways that complicate the management of their digital outputs. Virtual communities are often project-based and exist only so long as they serve the community or project that gave rise to them. Examples of virtual communities in the sciences are iemHUB-Integrated Environmental Modeling (http://iemhub.org/), CLEERhub: Collaboratory for Engineering Education Research (http://cleerhub.org/), the Mollsuc Health Laboratory' site at the University of Prince Edward Island-VRE-MHL-α (http://discoveryspace.upei.ca/mhl/), and CARMEN, a neuroscience community in the UK (http://www.carmen.org.uk/).

Software tools designed specifically for research-oriented virtual communities have begun to appear in the past few years. Some of the leading examples are HUBzero (http://hubzero.org), myExperiment (http://myexperiment.org), and Islandora (http://islandora.ca). Other discovery tools that support virtual communities, such as VIVO (http://vivoweb.org/), are appearing as well. Given the rise of such tools, research libraries should study them and the communities who are using them. They should begin to think

about and experiment with roles they can play in these spaces. To date, there appears to be only a small amount of organized, strategic activity from research libraries in virtual communities.

Recently, writers and organizations focusing on emerging roles for research libraries, such as Cliff Lynch (2008), Rick Luce (2008), and the Canadian Association of Research Libraries (CARL, 2010) have all described the possibilities of research librarians participating in the information enterprise that takes place in virtual communities.[13] Potential roles put forth have ranged from being RSS channel editors who package and deliver specific content to the virtual community to being research data managers who conduct the initial data collating and applying of metadata. Other roles include serving as information architects and designers who build and maintain the virtual community site, as communication facilitators who help connect people together in the community, and as information specialists who "lurk" on discussion forums and help connect people with the information they are seeking, be it a previous post to a forum, blog, or wiki, or a more formal publication. Moreover, research librarians and related information professionals can take the lead as virtual community managers and designers. They can facilitate research communication and information-finding in these virtual communities, speeding up the way researchers locate information shared in previous online discussions as well as the various information products they create. Research librarians can also curate these information products by identifying, describing, and managing them in the original virtual community site and/or in a digital repository that serves as a secure location for longer-term access and preservation purposes.

Because virtual communities exist to bridge institutional and sometimes national boundaries, policy decisions regarding who controls or has rights to the intellectual assets they generate can be an exceedingly complicated issue. Data sets, technical reports, and scholarly conversations documented via blogs, wikis, discussion lists, etc., all have intellectual value. Decisions have to be made as to who controls them, who can access them, who can share them, and who can preserve them. Rights management issues are another bailiwick in which research librarians can perform vital services for the research community to avoid such debilitating questions about rights and ownership, so that researchers can produce information as well as use and share it in a timely and effective fashion.

Endnotes

1 Anna Gold, "Cyberinfrastructure, Data, and Libraries, Part 2: Libraries and the Data Challenge: Roles and Actions for Libraries," *D-Lib Magazine* 13, no. 9/10 (September/October 2007). http://www.dlib.org/dlib/september07/gold/09gold-pt2.html.

2 This development not only applies to the sciences and engineering, but can be seen playing out in the digital humanities as well. In similar fashion, the roles of research libraries and librarians in the digital humanities are emerging from the process of data/information/content generation, as well as the related communications that occur in virtual settings and the need to sustain access to products of these scholarly and research processes.

3 The following reports are examples of studies that have started to document the formation of this data lifecycle management role:
Catherine Soehner, Catherine Steeves, and J. Ward. "E-Science and Data Support Services: A Study of ARL Member Institutions." Washington, DC: Association of Research Libraries (August 2010), http://www.arl.org/bm~doc/escience_report2010.pdf.

"Cyberinfrastructure Vision for 21st Century Discovery," Washington, D.C.: National Science Foundation (2007), http://www.nsf.gov/pubs/2007/nsf0728/index.jsp.
"To Stand the Test of Time: Long-term Stewardship of Digital Datasets in Science and Engineering," Washington, D.C.: Association of Research Libraries (2006), http://www.arl.org/pp/access/nsfworkshop.shtml.
"Long-Lived Digital Data Collections Enabling Research and Education in the 21st Century," Washington, D.C.: National Science Board (2005), http://www.nsf.gov/pubs/2005/nsb0540/.
D. E. Atkins, et al. "Revolutionizing Science and Engineering Through Cyberinfrastructure," National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure (January 2003), http://www.nsf.gov/cise/sci/reports/atkins.pdf.

4 This term was coined by information professionals at the University of Kansas: Richard Fyffe, D. Ludwig, and B. F. Warner, "Digital Preservation: A Campus-wide Perspective," *ECAR Research Bulletin* 18 (*EDUCAUSE*, August 2005), http://net.educause.edu/ir/library/pdf/ERB0518.pdf.
—"Digital Preservation in Action: Toward A Campus-wide Program," *ECAR Research Bulletin* 19 (*EDUCAUSE,* September 2005), http://net.educause.edu/ir/library/pdf/ERB0519.pdf.

5 Clifford Lynch, "The Institutional Challenges of Cyberinfrastructure and E-Research," Educause Review 43 (November/December 2008) http://connect.educause.edu/Library/EDUCAUSE+Review/TheInstitutionalChallenge/47446.

6 See: Brian E. C. Schottlaender and R.H. McDonald, "Data Cyberinfrastructure Collaboration at the University of California, San Diego," Coalition for Networked Information Task Force Meeting, Washington, DC (Fall 2007), http://chronopolis.sdsc.edu/publications.html. See other related reports and presentations at: http://chronopolis.sdsc.edu/assets/docs/cni_fall07.pdf.

7 See the Digital Curation Centre's digital lifecycle model for more information: http://www.dcc.ac.uk/resources/curation-lifecycle-model.
See also N. Beagrie, "Digital Curation for Science, Digital Libraries, and Individuals," *International Journal of Digital Curation* 1 (2006), http://ijdc.net/index.php/ijdc/article/view/6/0.
See also P. Manjula and A. Ball, "Challenges and Issues Relating to the Use of Representation Information for the Digital Curation of Crystallography and Engineering Data," *International Journal of Digital Curation* 1 (2006), http://ijdc.net/index.php/ijdc/article/view/64.

8 Michael Witt, et.al., "Constructing Data Curation Profiles," *International Journal of Digital Curation* 4 no. 3 (2009), http://ijdc.net/index.php/ijdc/article/view/137/0. Also see http://datacurationprofiles.org.

9 Catherine Soehner, Catherine Steeves, and J. Ward, "E-Science and Data Support Services: A Study of ARL Member Institutions," Washington, DC: Association of Research Libraries (August 2010), http://www.arl.org/bm~doc/escience_report2010.pdf.

10 For more information, see Purdue's Distributed Digital Curation Center (http://d2c2.lib.purdue.edu/), Johns Hopkins's Institute of Data Intensive Engineering and Science (http://idies.jhu.edu), Washington's eScience Institute (http://escience.washington.edu/), and Illinois's Center for informatics Research in Science and Scholarship (http://cirss.lis.illinois.edu/).

11 See Cornell's DISCOVER Research Services Group (http://drsg.cac.cornell.edu/) and the Library's DataStaR service (http://datastar.mannlib.cornell.edu/). For an example of MIT's service, see: http://libraries.mit.edu/guides/subjects/data/archiving/index.html.

12 See Johns Hopkins University's Data Conservancy (http://dataconservancy.org/home) and the University of New Mexico's DataONE initiative (https://www.dataone.org/) for more information.

13 Clifford Lynch, "The Institutional Challenges of Cyberinfrastructure and E-Research," *EDUCAUSE Review* 43, no. 6 (2008).
Richard E. Luce, "A New Value Equation Challenge: The Emergence of eResearch and Roles for Research Libraries," *No Brief Candle: Reconceiving Research Libraries for the 21st Century* (CLIR, August 2008).
Kathleen Shearer and Diego Argáez, "Addressing the Research Data Gap: A Review of Novel Services for Libraries," CARL: Ottawa, Ontario, Canada (March 2010).

Appendix B

Disciplinary Considerations for Digital Curation: The Digital Humanities

Digital Humanities Cyberinfrastructure and Data Curation

Digital humanities is an expanding field of inquiry that takes a methods-based approach to conducting humanities research in electronic forms. There are multiple facets or focuses of activity to this field, including both the praxis-based creation of digital scholarship on digital humanities topics (e.g., edited texts, digital archives, learning environments, and exhibits), and study of the impact media changes have on the scholarly enterprise (especially disciplinary inquiry) and on the world in which we live. The digital humanities are highly interdisciplinary and also incorporate a number of lenses and tools, from text encoding tagging to social media tools, GIS mapping, and multimedia gaming. Thus, digital humanities provide researchers with new paradigms through which they can approach long-standing cultural questions and issues. The emerging digital humanities community emphasizes the use of open standards and open source solutions and has, from its inception, entered into strong partnerships with research libraries. By carefully examining some of the recent developments in this field, we can begin to chart the potential roles libraries might play in future work in this area.

The evolution of the field of digital humanities and its recent maturation can be mapped, in part, by the ways in which these projects have been funded and supported. Many of the early projects of the last three decades have been created through foundation and federal grant support, usually through existing (i.e., non-digital in focus) grant programs. These grants usually required some level of university cost sharing (especially for the federal grants), so the universities have also provided some degree of support for this work throughout this time.

Over the last few years, an important transition has taken place, with major agencies like IMLS and NEH explicitly creating programs dedicated to support digital humanities projects. This is one of many markers that the field of digital humanities is solidifying. Such funding investments will continue to help institutions offset the costs of experimental work in this area. However, institutions are increasingly expected to provide certain parts of the cyberinfrastructure needed to support these endeavors in the long term. Among the top requirements (now a part of the grant application process for both NEH and IMLS applications) is that of preservation and long-term sustainability. Before grant agencies fund new humanities publications in the digital arena, they want assurance that the recipient has both the means and the university's commitment to preserve and make accessible that content for the long term. This component of the application process is helping to ensure that an appropriate relationship between the creators and the curators of digital content is established before the project begins. It also makes clear to the university and to the professors who pursue grant funding that the library has a role to play in these initiatives.

This is a promising development, and one that research libraries must capitalize on in order to serve as curators of these digital creations. Even still, many of the earlier trends in digital humanities work have *not* established this close connection between libraries and researchers. In many cases, research centers have instead been founded to take on this role. Such compartmentalization does not serve the research campus well as a whole; nor does it allow libraries to fulfill their duties as curators for their campus in the digital

medium. As the field continues to solidify, it is imperative that libraries determine what their campuses need in terms of digital humanities curation and that they quickly seek to meet those needs.

The types of projects included under the umbrella term of "digital humanities" range from text encoding initiatives (Brown University's Women Writers Project) to digital archives (UVA's Valley of the Shadow), and from e-journals (USC's Vectors) to full-scale learning environments (Emory University's Transatlantic Slave Trade "Voyages"). It also includes initiatives that have worked on curatorial issues, such as managing institutional history or federating collections from many distinct archives, and communications issues, such as the H-NET listserv and newer social networking tools. What are the implications of these different content types for research campuses, and what roles might we as libraries play as such content continues to be produced?

Creation of Digital Humanities Scholarship: Facilitating Sustainable Efforts

Some of the most promising digital humanities initiatives that we can point to in the still-emerging field are those that have effectively drawn together scholars, librarians, and information technologists to produce works that would not be possible without the perspectives and skill sets that each of these groups bring. Some of these ventures have occurred within existing structures, including the research library; others have happened within academic departments or through the creation of new centers that are focused on digital initiatives. Consider for example the Valley of the Shadow project (http://valley.lib.virginia.edu) at the University of Virginia and the blended team of scholars, graduate students, librarians, archivists, and technologists that brought this resource into being. The strong partnership that undergirded the development of this early resource enabled it to cross disciplines and genres, effectively becoming a major pioneer in the humanities computing landscape.

Libraries have partnered with humanities researchers to produce many of the earliest digital publications. Examples of such partnerships can be seen in the Text Encoding Initiative (TEI) centers that many of us have housed for well over a decade, formally or informally, and in the digital journals that we increasingly host, sometimes even as co-editors. We also see exemplar partnerships in the training offerings hosted by libraries for faculty and students, including lectures, workshops, courses, fellowships, and even graduate certificate programs. There is a broad range of experience across ARL libraries, though, and not every campus library is equally involved in the digital humanities infrastructure that continues to emerge on campuses nationwide.

In some instances, libraries have either chosen not to get involved or were not invited to participate in early initiatives. For example, faculty members have founded research centers devoted to a specialty area or topic, and digital humanities activities have flourished therein, without relying on the library for support. In other cases, units grew up in the library and then diverged pathways, eventually founding centers that provided these groups with a separate support infrastructure. As Diane Zorich cautioned, the structure of research centers, which typically operate as independent entities within the campus setting, does not effectively leverage campus-based resources. These centers often depend upon various campus agents in order to conduct their work (e.g., space donated from one campus entity, time and effort donated from others) in ways that are both difficult to quantify and difficult to replace if and when an agreement ends.[1] This has left some digital humanities centers vulnerable as targets for budget cuts in the current fiscal crises that most colleges and universities are experiencing, due to their perceived overlap in agenda and activities with other entities.

As a result, libraries have renewed opportunities to provide the local space and services that their institutional constituents need to continue their digital humanities development, as well as to cultivate the inter-institutional space and services that the field needs in order to thrive. Libraries offer a much-needed

central and well-connected meeting point for voices campus-wide who are involved in successful digital humanities programs.

As the digital humanities field continues its development, it is beginning to form standardized and normalized practices. As is true in many industries and fields in this time of rapid change, research libraries have a window of opportunity to define their roles. This opportunity must be used wisely: changes are already beginning to slow, and the transitions research institutions are making into the digital arena will continue to concretize. It will become increasingly difficult to challenge or change the roles research libraries choose for themselves today (by will or by default) in the future.

Digital Humanities Content: Needs and Expectations

Digital Humanities practitioners are producing many different types of content, all of which require ongoing management in order to maintain their viability. Regardless of the location of the digital humanities group(s) within the campus setting, the library has roles to play in helping these initiatives to produce and to manage sustainable resources for present and future generations.

Some of the potential roles for information professionals in managing and preserving content in the digital humanities can be summarized as follows:

1. Creation of humanities scholarship
    - Production of edited texts, digital archives, learning environments, journals, exhibits, and other e-pubs, including blogs and e-books
2. Curation of data and information objects
    - Managing and preserving increasingly complex digital objects and digital platforms, including interactive virtual communities

Creation of Humanities Scholarship

Historically, libraries have not served as producers of scholarship. Societies, university presses, and corporate entities most often provided this service to scholars in the world of the printed page. In contrast, in the digital environment, many libraries have engaged with scholars, often in research center settings, to help produce a wide variety of publishing ventures. These range from edited texts (e.g., the Women Writers Project, http://www.wwp.brown.edu; the Walt Whitman Archive, http://www.whitmanarchive.org; the Perseus Digital Library, http://ww.perseus.tufts.edu) to portals/archives containing diverse data types (e.g., the September 11 Digital Archive, http://911digitalarchive.org; the Valley of the Shadow, http://valley.lib.virginia.edu/), and from learning environments (e.g., Transatlantic Slave Trade Database Online, http://slavevoyages.org) to e-journals, e-books, and blogs (e.g., Vectors, http://www.vectorsjournal.org; Southern Spaces, http://southernspaces.org).

Librarians have become involved with these initiatives (and sometimes have helped to lead them) in order to assist scholars in activities that demand information management knowledge and expertise. Perhaps the most critical of these are the need for strategic information processing, federation of resources, digitization assistance, and increasingly, preservation expertise. This set of needs has opened up a potential growth space for the library, one that can serve the digital needs of scholars and researchers through in-house collaboration. Many of the most renowned digital humanities centers already operate in close collaborations with university libraries (e.g., MITH at Maryland, IATH at the University of Virginia, and the new Center for Digital Scholarship at Brown), and these engagements may help to provide a "renewed sense of library as laboratory as well as a physical and digital repository," as Patrik Svensson has recently noted.[2]

A natural alliance between scholars and archivists has helped to guide the development of many of the early digital library and digital archive resources. This has come primarily from the librarian-archivists'

relationship to the materials with which scholars have sought to work and the innovations demonstrated in many libraries by a "digital library" cluster of activity. As attention continues to shift toward the maintenance and sustainability of these digital creations, the library's role arguably becomes more prominent and essential. The often-heard phrase "preservation begins at creation" is key; bringing preservation expertise to bear on the creation of new resources (including how these are technically structured, what formats they use, and what metadata is created/extracted as the resource is developed) may be among the highest-value services that scholars need and libraries can provide in this still-emerging environment.

One of the challenges libraries face as they establish their place within the infrastructure of the digital humanities (intentionally or not) is that of reaching scholars with their services. Most libraries are still unaware of much of the digital humanities activity that is occurring on their campuses, whether this takes the form of what Cathy Davidson marks as Humanities 1.0 (first-generation and data-based projects) or Humanities 2.0 (more interactive scholarship marked by "a different set of theoretical premises, which decenter knowledge and authority,").[3]

Libraries will only be integrated into this work when the scholars who are undertaking it know that they are there. For example, a library may not know that several scholars on campus are producing blogs and e-books of note. Likewise, those scholars may not know that the library could assist them in publicizing (through library catalogs and digital library registries, for example) and sustaining their scholarly creations (through ensuring that production and, when necessary, migration of the resource(s) are conducted in ways that are consistent with broader practices in the field), and also through at least backing up, if not preserving, the content.

The danger in libraries not knowing what scholars are producing, and scholars not knowing what services libraries can provide in the digital humanities, is twofold. First, it may yield the loss of content that could and should have been sustained through a robust infrastructure. Already, many blogs, research databases, e-journals, and other "self-published" creations by scholars, which held significance within their disciplinary fields, have disappeared due to computer crashes, mismanagement, and natural disasters. By providing either back-up services (to ensure the persistence of the bits) or preservation services (to ensure the accurate renderability of the content over time), libraries can prevent such losses. Second, when scholars create new works without engaging with information specialists, they may miss opportunities to adhere to good practices in data structuring and data management. As a result, the sustainability of these resources will likely be more difficult down the road. Either way, the library incurs expenses downstream from creation (either through object loss or through acquiring the content at a stage when its structure makes it more costly to preserve). By engaging actively with the scholarly community in an ongoing way, librarians can avoid these pitfalls. Regularly surveying the community, using existing connections (e.g., subject liaisons), reaching out through other campus constituents who are likely to come into contact with digitally oriented scholars (offices of sponsored research, deans of faculty), and effectively marketing their services across departments may help. As with the sciences, libraries must expect that they will need to meet scholars where they are, not expect scholars to come to them.

Libraries are (rightly) concerned about accepting new responsibilities without simultaneously gaining new positions to offset that workload. However, libraries must also be concerned about losing ground within their campus environment by not meeting the digital needs of the scholarly community. If we do not seek to engage with the digital humanities, other entities will. Resources will follow that work, whether they are delivered by other campus-based groups or outsourced to external companies. Research libraries are in a pivotal moment of field formation in which they have the opportunity to expand their services in a way that is in keeping with their mission or to place boundaries around their staff and services that may relegate them to being bystanders in the digital arena. Being actively engaged and involved during this conceptual phase of digital humanities work will benefit both research libraries and the campuses they serve in the long term.

Curation of Data and Information Objects

Digital humanities projects and programs both within and beyond the library have now established a wide variety of collections for which libraries are increasingly providing (or being asked to provide) long-term management. This trend will likely escalate in the near future, as most federal granting agencies, which foster many of these projects and programs, now require that grantees contend with the issues of sustaining the works that they create. It is no longer enough to create a new resource for the field (e.g., a digital archive) or to produce scholarship in a web-based environment (e.g., an e-journal); as scholars and librarians undertake the creation of new digital forms of communication, they must also establish the means of ensuring its survivability.

As libraries continue to establish their role in the ongoing curation of digital data and information objects, questions will continue to arise around the responsibilities and costs associated with the management of these humanities collections and objects. What services should the library provide to its campus? How can these become ongoing components within the library's budget, where monitoring the 1s and 0s of digital objects and the server space they require is parallel to the monitoring of physical books and the shelf space that they, too, require? Who should bear the costs of curating and preserving these new collections?

Recent studies, including those driven by the Blue Ribbon Task Force, have demonstrated the complications of establishing the costs of providing long-term access to digital assets.[4] These studies have also shown that ingest is currently the most expensive part of the process.[5] With unknown long-term costs, most library deans are wondering what services can their libraries commit to provide for the digital humanities—and should they make any commitment at all?

We wonder if libraries need to reverse this question. What will happen if libraries refuse to make a commitment to curate and preserve digital resources for their campuses? Again, the mission of the library has long been to acquire, preserve, and provide access to research materials. In the academic research library setting, this is done primarily on behalf of the campus community, with a focus on the academic strengths of that campus. Can the library afford to not serve the community's needs in the digital environment, especially given that most scholarly production will likely be digital in the near future? If libraries wait for funding to begin addressing this core area in a systematic fashion, the field will find other locations for the hosting of its collections, whether those are on or beyond the institutional campus.

The question, then, might better be: how can libraries fulfill their missions in this time of change? One key strategy is to use every available mechanism to raise awareness within upper administrations of the new burdens that libraries are carrying on behalf of their campuses—including the faculty to whom libraries serve. Engaging humanities professors in this work, including having them help to perform gap analyses and peer comparisons, will help to elevate the necessary level of discussion at the campus level.

Educating scholars (and through them, the upper administration) about the costs of preserving digital humanities publications requires that research libraries make clear the value of their services. As we curate content, we also must provide scholars with guidance regarding the longevity of different formats and structures, back-up and preservation strategies (including migration), and the need for a better infrastructure within the library to ensure the longevity of the campus's scholarly endeavors.

We also must ensure that we make good, long-term decisions regarding how we manage the process of preservation on behalf of our campus communities. For what collections will best be served by centralized approaches to preservation, whether those are hosted locally, at the state level, or by an external provider? For what collections are better served by distributed preservation models and shared cyberinfrastructures across many institutions that hold a shared interest in those materials? Should we use different mechanisms for the preservation of ETDs versus e-journals? Should we turn to different partners as we preserve digitized book collections versus specialized digital archival creations? And what role should we play and what mechanisms should we employ in the preservation and maintenance of web-based materials that are

created well beyond the boundaries of research campuses, but will be necessary primary resources for tomorrow's scholars?

These questions require study, and the digital humanities is one key arena in which research libraries' development activities and field-based knowledge have matured to a degree to allow for some of that study to take place. Capitalizing on that potential knowledge, rather than resisting the extra work it will require, may be in the libraries' long-term interests as institutions.

For example, we can now study the evolution of resources (and the associated costs of that evolution) built more than a decade ago through digital humanities ventures, such as early TEI projects and the digital archives "Valley of the Shadow" collection (UVA). We can evaluate the sunk costs, grant funding, and outright expenditures that contributed to the maintenance of these resources to establish one baseline for the costs associated with the long-term management of digital resources. We can even begin to compare those costs to the anticipated maintenance costs associated with some of the increasingly complex multimedia webs of the "digital humanities 2.0" landscape, including the Voyages and Origins sites (http://slavevoyages.org and http://www.african-origins.org), which incorporate a wide range of media types, as well as researcher and public contribution pathways that are overseen by editorial boards.

We can also use such examples to begin to discuss preservation cost models that better reflect the collaborative investment that happens in the development cycle of many of the digital humanities projects and publications. For example, the Voyages and Origins sites were both grounded in data gathered by five main scholars (and colleagues, graduate students, and post-docs along the way) who were located across multiple continents. Although all five scholars and their institutional infrastructures contributed in various ways to the creation of the datasets, and many more in the extended scholarly and librarian communities that helped with the development of additional resources for the web-based publications, the sustainability plan currently depends on the lead scholar who undertook the responsibility for developing the sites and the willingness of his institution to maintain them into the foreseeable future. Placing the burden of maintenance on one scholar or one institution is not likely to be a viable way to structure preservation for these resources in the future. A key question is how to engage the extended group that helped to put a digital humanities project or publication together—and, of course, its user community when possible—to help to ensure its longevity. What types of co-curation and content management might work across institutional boundaries? We continue to need experimentation with new models for these types of interactions and engagements.

Endnotes

1 Diane Zorich, *A Survey of Digital Cultural Heritage Initiatives and Their Sustainability Concerns*, (Washington, DC: CLIR, 2003), http://www.clir.org/pubs/reports/pub118/contents.html.

2 Patrik Svensson, "The Landscape of Digital Humanities," *Digital Humanities Quarterly* 4, no. 1 (2010), http://digitalhumanities.org/dhq/vol/4/1/000080/000080.html.

3 Cathy Davidson, "Humanities 2.0: Promise, Perils, Predictions," *Publications of the Modern Language Association of America* 123, no. 3 (2008): 712.

4 For more information, see "Sustainable Economics for a Digital Planet-Blue Ribbon Task Force," http://brtf.sdsc.edu/biblio/BRTF_Final_Report.pdf.

5 For more information, see "Keeping Research Data Safe (Phase 2)," which is available at http://www.jisc.ac.uk/publications/reports/2010/keepingresearchdatasafe2.aspx#downloads.