Move It or Lose It: Investigating Digital Curation Portability for Access to Government Information

Christopher A. Lee; University of North Carolina; Chapel Hill, North Carolina, USA

Abstract

A fundamental issue of digital preservation is that information resources must often out-live the systems that are used to maintain them at any given time. It is also important to consider sustainability across the boundaries of collection environments. Portability is an essential consideration. The project called "A Model Technological and Social Architecture for the Preservation of State Government Digital Information," administered by the Minnesota Historical Society, is developing strategies and systems to provide enhanced online access to state legislative materials. The project is testing software and strategies to collect and provide access to state legislative documents and associated contextual information. The long-term sustainability of the effort will require interoperability among a various parties, including (1) those who might share responsibility for the preservation of legislative resources from Minnesota, and (2) collecting institutions from other states who would like to make use of the project's methods and software. The author is investigating characteristics of the state legislative information system that are most likely to support or hinder portability of software and digital objects across the boundaries of organizations. The findings from this investigation should be relevant to information professionals responsible for digital collections or collection management systems that must be sustained across the boundaries of specific technical or organizational arrangements.

Introduction and Motivation

A fundamental issue of digital curation is that information resources must often out-live the systems that are used to maintain them. Within a given repository, this issue is addressed, in part, by the Preservation Planning function of the Reference Model for an Open Archival Information System (OAIS) [1], by ensuring that information persists across changes in hardware and software. It is also important to consider sustainability across the boundaries of archives. Archives can enter relationships in which commitments to resources are shared; the OAIS identifies cooperating, federated and shared resource associations.

Over time, sustainability can also require "an appropriate, formal succession plan, contingency plans, and/or escrow arrangements in place in case the repository ceases to operate or the governing or funding institution substantially changes its scope" [2]. The literature related to digital archives has generally treated transfer in such scenarios as a relatively discrete event, in which entire collections are moved from one environment to another. A prominent example is the Archive Ingest and Handling Test (AIHT), which tested the "wholesale transfer, ingestion, management, and export of a relatively modest digital archive,

whose content and form are both static" [3]. Despite the inter-institutional structure of the transfer scenarios, the repositories involved in the AIHT "chose to place their trust in their own local tools and processes" rather than incorporating outputs of tasks performed by the other repositories into their own ingest processes [4].

Investigations and documentation of the requirements for transferring static sets of content across institutions are both important and valuable. However, long-term curation of digital content may involve "a series of handoffs, occurring repeatedly at many levels: between different types of media and storage subsystems, different object frameworks and organizational schemes, different repository systems, different institutions and policy regimes, and different application communities with diverse assumptions and interests" [5].

Arrangements for inter-archive coordination and inter-archive sustainability of collections raise legal, ethical, economic and organizational issues. They also raise numerous technical questions. Archives are not simply aggregates of digital objects; they are also composed of functions, services and internal relationships that are supported and enabled by software. When moving information resources across archive boundaries, portability is an essential consideration. In the digital preservation literature, there has been relatively little investigation of characteristics that support the re-use of information systems across the boundaries of archives.

This paper discusses information systems for the curation of state legislative information, including system characteristics and strategies that are most likely to support or hinder portability of both software and digital objects across the boundaries of organizations. The findings from this investigation should be relevant to information professionals who are responsible for digital collections or collection management systems that must be sustained across the boundaries of specific technical or organizational arrangements.

Legislative Records as Online Public Information

For thousands of years, records have served as instruments of authority and power [6]. The tradition of state archives providing public access to records is relatively new, dating back to shortly after the French Revolution. Even more recent is the phenomenon of governments providing widespread access to information about their activities to citizens through the Internet. More recent still have been systematic efforts to facilitate public discovery and retrieval of online government information; notable efforts at the federal level in the U.S. have included the Government Information Locator Service (GILS) effort initiated in 1993, THOMAS in 1995, passage of the Electronic Freedom of

Information Act Amendment (EFOIA) in 1996, and the launch of FirstGov.gov (now USA.gov) in 2000, Regulations.gov in 2002, and Data.gov in 2009. Entities from outside the government have also established online resources that citizens can use to hold public officials accountable for their statements and actions, prominent examples in the U.S. being FactCheck.org, PolitiFact.com, and the National Security Archive.

Online access to public information has the potential to advance both the efficacy and legitimacy of government, by enhancing service offerings and allowing citizens to actively engage in governance processes. An essential piece of this puzzle is state legislative information, which documents current and previous policies, as well as the drafting and approval process. Online access to legislative information can support new forms of investigation that are afforded by digital data, particularly if one can search and analyze materials from multiple states.

State Legislative Records Projects

In 2005-2008, the Minnesota Historical Society (MHS), Minnesota Office of the Revisor of Statutes (ROS), and Minnesota Legislative Reference Library (LRL) engaged in a project funded by the National Historical Publications and Records Commission (NHPRC) called "Preserving the Records of the E-Legislature," which aimed to explore and test the technologies available to preserve the electronic records of the Minnesota legislature. The project began the same year that the ROS implemented a new drafting and document management system called XTEND, which is discussed below. The partners received technological guidance and services from the San Diego Supercomputer Center. The California State Archives, State Library, and Legislative Counsel also provided input and considered applicability of the project's product to the California context. The project generated documentation of the ROS workflow and document types; articulated options for further development efforts; and provided a solid foundation for further collaboration. One of the conclusions was that "the essential records to acquire and preserve [from the ROS] are session laws, statutes, and administrative rules, all of which can be relatively easily extracted from an XML-based bill drafting system [XTEND] at the end of a session" [7].

A project now underway, called "A Model Technological and Social Architecture for the Preservation of State Government Digital Information," is developing strategies and systems to provide enhanced online access to state legislative materials. It is led by MHS, and funded by the National Digital Information Infrastructure and Preservation Program (NDIIPP) of the U.S. Library of Congress. Partners include the Minnesota ROS, Minnesota Legislative Reference Library, University of California Curation Center (UC3), California's State Library and State Archives, California Legislative Counsel, and National Conference of State Legislatures. There are many other participating states, including Arkansas, Illinois, Kansas, Mississippi, Nebraska, North Dakota, Tennessee, and Vermont.

The project is testing software and strategies to collect and provide access to state legislative documents and associated contextual information. In the short term, one of the main system requirements is interoperability between the MHS and UC3 technical environments and the primary sources of legislative information in the state, most notably the Minnesota ROS and Legislative Reference Library.

Longer-term sustainability of the effort will require interoperability among a larger set of organizations, including (1) those who might share responsibility for the preservation of the legislative resources from Minnesota, and (2) collecting institutions from other states who would like to make use of the project's methods and software.

MHS Approach to Electronic Records

The current project grows out of an ongoing program of the MHS to engage and collaborate with state entities in Minnesota and other states toward the goal of better managing and providing access to digital assets. The MHS approach since the 1990s "can be characterized by close collaboration with government constituents, the development of practical tools, and an emphasis on education" [8]. A core element has been the idea of "trustworthy information system," which an agency defines for itself, based on its own constraints, responsibilities and priorities. MHS has offered considerable support, education and guidance to agencies in this process, including the development of the *Trustworthy Information Systems (TIS) Handbook* [9]. Further educational efforts have included a multi-state project funded by the NHPRC to develop workshops on XML and metadata.

Minnesota Revisor of Statutes System (XTEND)

The Minnesota ROS uses a legislative document processing system, called XTEND (Xml Text Editor, New Development). The XTEND system includes a storage area network, database, application server, web server, text editing application and publishing engine. Data passed between the components of the system is a combination of XML, relational data, and Java objects. The document editing and publication components deal with documents encoded in XML.

Since going into production in 2005, XTEND has provided direct public access to a variety of document types, disseminated in PDF (Portable Document Format) or XHTML (Extensible Hypertext Markup Language). In March 2006, the ROS staff carried out an analysis of XTEND based on the *TIS Handbook*. The ROS has also recognized and articulated limitations of XTEND in supporting emerging discovery, retrieval and aggregation scenarios [10].

Prototype System for Integrated Access

One of the goals of the collaboration between the ROS and MHS is to better support integrated search and retrieval of legislative documents by taking further advantage of the XML encoding of the documents. Two components of this effort have been the development of a schema for XML wrappers of content to be transferred; and a prototype content management and search environment based on an XML-native database.

An XML Schema Working Group -- including representatives from the MHS, ROS, XMaLpha Technologies, and Thomson Reuters – has proposed a packaging approach for the transfer of legislative data [11]. It allows for four main elements:

    <metadata> - to convey to any supplementary descriptive, technical or administrative metadata that it has to offer.

    <xml.source> - XML version of a bill

<html.rendition> - HTML version of a bill
<attachments> - any versions of a bill besides XML or
    HTML (e.g. PDF, Word).

A prototype system (written in PHP) has been developed to generate XML wrappers for ROS legislative documents. If a user enters the URL for a package into a web browser, she can then download a .zip file that contains three files containing: XML data, a SHA1 checksum and a MD5 checksum. The checksums can be used to ensure that the file transfer was successful. Opening the XML file reveals the metadata, XML, HTML, and PDF instances of the document.

The project has also been exploring the development of software that can better take advantage of the XML encoding of the legislative documents, in order to facilitate discovery and access. MHS has contracted with a company called Syntactica to develop a proof-of-concept software suite. The suite includes 17 small, integrated applications, built on top of eXist, which is an XML-native database [12]. The software was designed to allow staff without significant programming experience to develop new search interfaces, aggregations and data re-use or data management features, through the use of XQuery and style sheets. The software has been tested with a set of approximately 6000 documents from California, Illinois, and Minnesota.

It is important to note that the category of applications that are commonly called "XML-native databases" do not actually use XML as their underlying data structure, i.e. when importing XML, they do not store the intact XML files. Their internal data structures are based on stripping out the data values from the XML and then reorganizing and re-encoding the data in ways that are optimized for search and reuse of the data. The differences between an XML-native database and relational database are based on how they logically organize the data. The XML-native database organizes data in a way that attempts to reflect the hierarchical structure and order of the data elements from an XML file (e.g. <p> is a child element of <body>, <chapter2> came after <chapter1> within the document). By contrast, a relational database organizes data into tables, records and fields; and the process of "shredding" an XML source file into the structures of a relational database can potentially lose some of the hierarchical and ordering relationships that were part of the original XML. Direct access at the bitstream level is not the intended means of providing interoperability across applications. Instead, agents (users or applications) are expected to issue queries to the database in order to get data out of it. Relational database management systems often have their own proprietary extensions and flavors of query language, but the primary industry standard for queries is Structured Query Language (SQL). Relational databases benefit from this very well-established query language, based on several decades of experience with these structures. XML databases are a much more recent development. It appears that XQuery, XPath and XSLT (Extensible Stylesheet Language Transformations) will serve as the industry standards for querying XML data, though there could still be significant evolution of the standards and their implementation in years to come.

The project team has made a strong case for the use of an XML-native database to provide access to the XML-encoded legislative documents [13]. Searching and navigating XML files can be greatly facilitated by a system that exploits the XML data

elements, rather than working through intermediate relational database structures. At a meeting of state partners on the project on January 20, 2010, Isaac Holmlund from the ROS provided examples in which eXist and XQuery can reduce the effort required to add further arbitrary search or navigational elements to a user interface. The prospects of providing search across XML content from multiple states are very compelling. There could also be great potential for exposing this interface to users in ways that allow them to define their own, novel access points.

As noted above, the fact that eXist is an XML-native database does not mean that it actually stores data as serialized XML, i.e. in the specific form in which it was received. Instead, data being creating, managed and accessed by the database software takes the form of binary files. Making direct use of these binary files is dependent on the eXist database application. The internal data storage of eXist is based on a set of binary files with a .dbx extension, which which are based on (B+ trees and paged files).

There are two reasons for MHS to also manage the XML data from ROS in its original form outside of eXist. First, as explained above, eXist uses a unique binary data format that is not directly readable by other software. The second reason is related to the integrity of the original XML files. eXist has various options for serializing the data that it stores, i.e. exporting the data back out as XML – either as a text file or SAX (Simple API for XML) file. However, the way eXist breaks up and stores the data from XML file could prevent it from being able to generate an exact (at the bitstream level) copy of the XML file. In other words, the XML input will often not be identical to the XML output.

This could potentially be addressed by "canonicalizing" [14] the XML after it is received and before it is imported into eXist. Canonicalizing an XML document is a process of transforming its content into a specified form that enforces consistency of factors such as line feeds, white space, attribute values, attribute order, character encoding, use of namespaces, and Uniform Resource Identifiers (URIs). Two disadvantages of such canonicalization would be 1) the requirement to create a "staging area" between submission and important into eXist and 2) if one were to rely solely on the canonical form and abandon the original XML as received, this would eliminate the possibility of verifying that the hash value of the managed bitstream is the same as the hash value of the bitstream as received.

The eXist system has significant promise as one element in the full repertoire of offerings to support access to government information over time. The MHS project team is actively pursuing complementary activities that address other aspects of the digital curation landscape. One of those activities is its collaboration with the UC3.

Secondary Preservation Environment (UC3)

MHS has entered an arrangement with the University of California Curation Center (UC3) to enable MHS to use the UC3 tools and infrastructure to transfer, ingest, store, and report on legislative content, as well as exploring use of UC3 as an off-site repository. This exploratory work will include content from both Minnesota and the California Legislative Counsel, including data collected through the UC3's Web Archiving Service.

The design philosophy driving UC3 is that "rather than relying on a conceptually monolithic system as a locus, curation

outcomes will be the product of loosely coupled, independent, distributed services" [15]. Rather than proposing one consistent content model for all collections, the UC3 architecture is based on a set of well-defined file and directory naming conventions that are applied at the filesystem level. The UC3 architecture currently identifies twelve discrete services (called micro-services) that are divided into four hierarchical service layers: protection, interpretation, application and interoperation. The UC3 is currently in the process of developing and integrating many of the identified micro-services. The UC3 team's goal is to develop a set of services that will be "flexible with regard to local policies and practices" as well as "the inevitability of disruptive change in technology and user expectation" [16].

Transfer Scenarios and Considerations

Repositories manage packages that contain files they've acquired, rather than only managing the files themselves. The packages can associate files to each other and can also contain various forms of descriptive, technical and administrative metadata that should be associated with the files. The Reference Model for an Open Archival Information System (OAIS) identifies three distinct types of packages: Submission Information Package (SIP), Archival Information Package (AIP) and Dissemination Information Package (DIP).

There are three different kinds of SIPs that MHS/UC3 is likely to receive from the ROS in this project:

1. xmlwrapper files – These would conform to the conventions from the ROS described above. The contents within the xmlwrapper may include an XML file, but it may also contain other types of files (e.g. HTML, Word, PDF) and some associated metadata.
2. XML files that are not wrapped within a larger XML package – This could include, for example, XHTML files captured from the Web, possibly embedded in WARC (Web ARChive) [17] file as output from a Web crawl.
3. Files in formats other than XML that are not wrapped within a larger XML package – This could include, for example, MS Word documents captured from the Web or obtained directly from government agencies.

There are many different options for how MHS and UC3 could approach the workflow for acquiring and managing packages from the ROS. A promising scenario could be transferring "raw" copies of all content selected for retention to the UC3 environment directly from ROS (in a process mediated and supported by MHS), while also transferring files to MHS for import into the eXist system to facilitate search and access. Such a parallel arrangement could take advantage of the complementary services offered by the two environments, as well as reducing reliance on MHS's network bandwidth, which has been a limited factor in previous investigations. For example, eXist could be well-suited to obtain submissions of type 1, pull out the individual content files, and then import them into its database. It could also be well-suited to importing submissions of type 2. The UC3 could then manage the received files (of type 1, 2 and 3) as bitstreams to be preserved in their original form. Note that neither type 1 nor type 2 would serve as the AIP within a repository. The AIP will include preservation (discussed below) and other administrative metadata

that will need to be stored outside of the received files but associated with them. Many of the AIP metadata elements will change over time, so they cannot simply be included in the xmlwrapper that is originally submitted. The feasibility of importing packages of type 3 into the eXist environment is still an open question subject to further investigation.

As explained above, one of the requirements of managing AIPs is the ability to periodically update metadata values within the package. eXist was not originally designed to deal with data elements that are added, removed, or changed, though there has been some work over the past couple years to address this limitation of the software. By contrast, the UC3 has identified detailed procedures for addressing the versioning of data over time [16]. This is an example of how the two different environments could offer complementary services.

Implications and Future Directions

This paper reports on work that is still very much in process. It raises a number of open questions, including the further definition of roles and workflow; collection and aggregation of contextual information; authentication of state publications; comparison and collaboration with other projects; evolution and sustainability; and movement from the interoperability of systems to the portability of digital curation.

Further Definition of Roles and Workflow

A digital repository can be seen as a combination of services, resources (required to carry out and supported by the services), and policies that determine how the services should be implemented. One of the fundamental design questions is how to break down the services and resources: who will have responsibility, where they will reside, and how they will interact [18]. The appropriate answer will depend on "context-specific attributes" including value, incentives, roles and responsibilities [19].

The previous project funded by the NHPRC articulated a potential workflow in which MHS would perform many archival management activities remotely, rather than serving as the primary repository environment for the legislative documents. There is value in further articulating how such an arrangement will be designed and implemented. For example, one of the roles that UC3 intends to play within the University of California system is "service brokerage to select appropriate curation service providers and mediation of service-level agreement negotiation" [15]. In the case of records from the state of Minnesota, MHS is exploring the provision of such a service brokerage role for state government entities, including the ROS.

One example of an issue to be addressed in a distributed environment is the implementation and documentation of integrity checks. MHS could set the policies and procedures for execution of integrity checks on the data that resides in the storage environments of one or more trusted third parties. MHS could then issue requests – based on a defined schedule or designated trigger events – to those storage environments to generate checksums on designated files, objects or collections, and then report the results to MHS. MHS could then serve as the primary steward of the data that documented the execution of the integrity checks. This service would complement the UC3, which has established mechanisms for generating and reporting checksums

but does not currently have the capacity to represent a policy to govern the frequency of fixity checking within its services [16].

Collection and Aggregation of Contextual Information

Making meaningful use and sense of the digital objects often requires capture, collection and management of contextual information [21]. The initial focus of the MHS NDIIPP project is a subset of the materials from the XTEND system that have clear retention value. XTEND also manages many documents involved in the revisions of rules and statutes that are not disseminated to the public and will not be transferred to an external repository. However, there are also many documents related to the legislature and its activities that can be archived from the Web. The capture and retention of associated web content can potentially provide a great deal of contextual information that would be too costly and burdensome for archival professionals to generate themselves [22]. The UC3 will use its Web Archiving Service to explore the collection of materials related to both the Minnesota and California legislatures.

Authentication of State Publications

An open question that reaches far beyond the scope of the current project is what, from a legal perspective, constitutes an authentic copy of a government publication. In the U.S., the National Conference of Commissioners on Uniform State Laws is in the process of developing a proposed model act that would identify the broad conditions under which an official publisher of legal material can designate an electronic version as the "official" version [23].

In the case of Minnesota legislative materials, the ROS uses a secure transfer protocol (HTTPS) and a digital certificate for serving its web pages. The UC3 has tools and conventions for ensuring the fixity of bitstreams over time. It will be important to investigate what further mechanisms and provisions should be in place to ensure and document the chain of custody of the materials. This will be an evolving discussion, as both the technical and legal landscape are likely to change dramatically in years to come.

A well-established set of requirements for the curation of digital collections are the creation, capture and management of metadata associated with the origins of data and transformations or actions upon the data over time. The most widely recognized source of guidance for digital preservation metadata, including provenance and chain of custody, is Preservation Metadata: Implementation Strategies (PREMIS) [24].

Comparison and Collaboration with Other Projects

The project reported in this paper is part of a much larger ecosystem of projects and digital curation activities. One category of particularly related projects are others supported by the NDIIPP program, including several that are focusing on the preservation of state government information [25]. Another related initiative is Kansas Enterprise Electronic Preservation (KEEP), which aims to develop an enterprise-wide system for the preservation of state records, enabled by recent legislation related to the "maintenance and certification of electronic records."

The project team has been actively engaging other parties who have an interest in the outcomes of the work. Related activities are also serving as points of comparison and sources of ideas and expertise.

Evolution and Sustainability

The collective endeavor of long-term curation of state legislative records can be characterized as a system of systems (SoS) with the shared purpose of ensuring perpetual access to and meaningful use of the records. There are two defining features of an SoS: (1) its components fulfill "valid purposes in their own right" and "continued to operate to fulfill those purposes if disassembled from the overall system, and (2) the components systems are managed (at least in part) for their own purposes rather than the purposes of the whole" [26]. In this case, components include the systems of the Minnesota ROS, MHS, and UC3, as well as numerous systems across other states that may feed into or draw from the data and services of the NDIIPP project. A common SoS problem is "failure to architect for robust collaboration when direct control is impossible" [26]. Rather than planning for a single, centralized system development process, participants should focus on "intermediate systems" that are "capable of operating and fulfilling useful purposes before full deployment or construction is achieved" [26]. As the SoS evolves, it should be in the "interests of each participant to continue to operate rather than disengage" [26]. SoS success stories tend to be those in which "systems incrementally developed and evolved with continual integration incorporating tests for interoperability issues as they are discovered" [27]. A major focus of a SoS is the interfaces between systems.

The long-term curation of digital content will be best served through "robust design" [28], which can serve short-term purposes but is also sufficiently flexible to remain effective in a wide range of possible future contexts. Rather than attempt to stringently predict and control the entire SoS, those with an interest in long-term curation can attempt to "harness" its complexity [29] in socially beneficial ways. For example, one can potentially pre-empt some of the complication of cross-institutional exchanges by focusing on interoperability early in the process [30].

From Interoperability to Portability

From an engineering perspective, interoperability is the ability for two or more systems or functional units to "exchange information and to use the information that has been exchanged" [31] or "communicate, execute programs, or transfer data…in a manner that requires the user to have little or no knowledge of the unique characteristics of those units" [32]. Interoperability can greatly facilitate coordination and communication across systems, but is not itself sufficient for the types of digital curation transfers discussed in this paper.

One of the challenges of coordinating digital curation work is that the interfaces must not only support the transfer of data (resources), but also services and policies. A significant portion of a repository's value, for example, can reside in the services that it provides [33]. Long-term digital preservation will require the transfer of data across systems, which can be seen as an issue of interoperability between the present and the future [34]. This is an ongoing effort.

The establishment of mechanisms for interoperability (e.g. data transfer protocols) between two institutions can be a necessary

condition for collaborative curation of data, but it is not a sufficient condition for success. The institutions must also share expertise, services, practices, expectations, norms and mechanisms for ensuring trust. In other words, they must move beyond planning for interoperability to planning for digital curation portability. Regardless of the specific institutional or architecture arrangements, digital curation will also require ongoing professional engagement and the ability to collectively respond to a changing environment.

Acknowledgements

References

[1] CCSDS, Reference Model for an Open Archival Information System (OAIS): Blue Book CCSDS 650.0-B-1 (2002).

[2] Trustworthy Repositories Audit & Certification: Criteria and Checklist (Center for Research Libraries, Chicago, IL, 2007).

[3] Clay Shirky, Library of Congress Archive Ingest and Handling Test (AIHT) Final Report, National Digital Information Infrastructure and Preservation Program (2005).

[4] Martha Anderson and Bill LeFurgy, The Archive Ingest and Handling Test: Implications of Diverse Content and Diverse Repository Practices, Proc. Digital Curation & Trusted Repositories: Seeking Success (2006).

[5] Greg Janée, James Frew and Terry Moore, Relay-Supporting Archives: Requirements and Progress, International Journal of Digital Curation, 1, 4, (2009) pg. 57-70.

[6] Randall C. Jimerson, Archives Power: Memory, Accountability and Social Justice (Society of American Archivists, Chicago, 2009).

[7] Robert Horton, Shawn Rounds and Elizabeth Lighthipe, Preserving the Records of the E-Legislature: Final Project Report (Minnesota Historical Society, St. Paul, MN, 2008).

[8] Robert Horton, Obstacles and Opportunities: A Strategic Approach to Electronic Records, in Effective Approaches for Managing Electronic Records and Archives (Scarecrow Press, Landham, MD, 2002) pg. 53-71.

[9] Shawn P. Rounds and Mary P. Klauda, Trustworthy Information Systems Handbook, Version 4 (Minnesota Historical Society, St. Paul, MN, 2002).

[10] Tim Orr, Preserving State Government Digital Information, NDIIPP Project Partners Meeting, December 8 (2008).

[11] Tim Orr and Isaac Holmlund, An 'Xmlwrapper' for the Exchange and Archive of Legislative Bills, National Association of Legislative Information Technology Newsletter (Winter 2010) pg. 9-12.

[12] Wolfgang Meier, eXist: An Open Source Native XML Database, in Web, Web-Services, and Database Systems (Springer, Berlin, 2003) pg. 169-83.

[13] Nancy Hoffman, XML Native Databases and Legislative Documents: A White Paper (Minnesota Historical Society, St. Paul, MN, 2009).

[14] Clifford Lynch, Authenticity and Integrity in the Digital Environment: An Exploratory Analysis of the Central Role of Trust, in Authenticity in a Digital Environment (Council on Library and Information Resources, Washington, DC, 2000) pg. 32-50.

[15] Stephen Abrams, Patricia Cruse & John Kunze, Preservation Is Not a Place, International Journal of Digital Curation, 4, 1 (2009) pg. 8-21.

[16] Stephen Abrams, John Kunze and David Loy, An Emergent Micro-Services Approach to Digital Curation Infrastructure, Proc. iPRES p.4-11 (2009).

[17] J. Kunze, A. Arvidson, G. Mohr and M. Stack, The WARC File Format (Version 0.9), IIPC Framework Working Group (2006).

[18] Barbara Sierman, Raymond Van Diessen and Christopher A. Lee, Component Business Model for Digital Repositories, Proc. iPRES (2008).

[19] Blue Ribbon Task Force on Sustainable Digital Preservation and Access, Sustainable Economics for a Digital Planet: Ensuring Long-Term Access to Digital Information (2010).

[20] Christopher A. Lee, Richard Marciano, et al, Mainstreaming Preservation through Slicing and Dicing of Digital Repositories: Investigating Alternative Service and Resource Options for *ContextMiner* Using Data Grid Technology, Proc. iPRES, pg. 113-20 (2009).

[21] Christopher A. Lee, A Framework for Contextual Information in Digital Collections, Journal of Documentation (Forthcoming).

[22] Christopher A. Lee and Helen R. Tibbo, Capturing the Moment: Strategies for Selection and Collection of Web-Based Resources to Document Important Social Phenomena, Proc. IS&T Archiving, pg. 300-305 (2008).

[23] National Conference of Commissioners on Uniform State Laws, Authentication and Preservation of State Electronic Legal Materials Act, Interim Draft (2010).

[24] PREMIS Data Dictionary for Preservation Metadata, Version 2.0 (2008).

[25] Preserving State Government Information, Library of Congress, http://www.digitalpreservation.gov/partners/states.html.

[26] Mark W. Maier, Architecting Principles for Systems-of-Systems, Systems Engineering, 1, 4 (1998) pg. 267-84.

[27] Carol A. Sledge, Reports from the Field on System of Systems Interoperability Challenges and Promising Approaches (Software Engineering Institute, Pittsburgh, PA, 2010).

[28] Andrew B. Hargadon and Yellowlees Douglas, When Innovations Meet Institutions: Edison and the Design of the Electric Light, Administrative Science Quarterly, 46, 3 (2001) pg. 476-501.

[29] Robert Axelrod and Michael D. Cohen, Harnessing Complexity: Organizational Implications of a Scientific Frontier (The Free Press, New York, NY, 1999).

[30] Andreas Aschenbrenner, Tobias Blanke, et al, The Future of Repositories? Patterns for (Cross-)Repository Architectures, D-Lib Magazine, 14, 11/12 (2008).

[31] IEEE Standard Computer Dictionary: A Compilation of IEEE Standard Computer Glossaries (Institute of Electrical and Electronics Engineers, New York, NY, 1990).

[32] ISO/IEC 2382-01, Information Technology Vocabulary, Fundamental Terms (1993).

[33] Robert Chavez, Gregory Crane, et al, Services Make the Repository, Journal of Digital Information, 8, (2007).

[34] Margaret Hedstrom, Exploring the Concept of Temporal Interoperability as a Framework for Digital Preservation, Proc. DELOS (2001).

Author Biography

Christopher (Cal) Lee is Assistant Professor at the School of Information and Library Science at the University of North Carolina, Chapel Hill. He teaches archival administration, records management, digital curation, and information technology for managing digital collections. His research focuses on long-term curation of digital collections and stewardship (by individuals and information professionals) of personal digital archives. He is particularly interested in the practical and ethical implications of diffusing tools and methods into professional practice.