The International Journal of Digital Curation
Issue 1, Volume 4 | 2009

Relay-supporting Archives: Requirements and Progress

Greg Janée, James Frew,
University of California, Santa Barbara
Terry Moore,
University of Tennessee, Knoxville

November 2008

Abstract

We characterize long-term preservation of digital content as an extended *relay* in time, in which repeated handoffs of information occur independently at every architectural layer: at the physical layer, where bits are handed off between storage systems; at the logical layer, where digital objects are handed off between repository systems; and at the administrative layer, where collections of objects and relationships are handed off between archives, curators, and institutions. We examine the support of current preservation technologies for these handoffs, note shortcomings, and argue that some modest improvements would result in a "relay-supporting" preservation infrastructure, one that provides a baseline level of preservation by mitigating the risk of fundamental information loss. Finally, we propose a series of tests to validate a relay-supporting infrastructure, including a second Archive Ingest and Handling Test (AIHT).

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. ISSN: 1746-8256 The IJDC is published by UKOLN at the University of Bath and is a publication of the Digital Curation Centre.

Introduction

Creating a viable architecture for the long-term preservation of digital content is widely recognized to be both urgent and difficult (Hedstrom et al, 2003; Ross & Hedstrom, 2005). It is urgent because the amount of information now being "born digital" is increasing exponentially. Since preserving such content is an active process, data objects that are not archived reasonably soon by methods designed for long-term sustainability (e.g., a century or more) are likely to be lost. But digital preservation is a complex, multilayered exercise in social cooperation extended over time. Non-trivial changes will inevitably occur in the social and technological systems that implement it, and various factors can cause the last copy of any given item to become permanently inaccessible (Rosenthal, Robertson, Lipkis, Reich & Morabito, 2005). Our experience with this process clearly suggests that preserving digital content for a century will require 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. The design of such an "archive relay" for digital content must focus on achieving the kind of interoperability that maximizes the ease with which such handoffs can successfully be made, in spite of the heterogeneity that will be introduced at many steps along the way (Hedstrom, 2001).

Traditional archive design focuses on archive *creation*, where some digital content, which has been produced recently enough to still be usable, is perceived to have significant future value, and action must be taken soon to avert its loss. However, if a collection archived today is to be available and usable a century from now, we must presume that it will migrate through a series of archives, possibly in distant locations, using different technologies and under different administrative controls. Preservation is more assured if the archives along that sequence interoperate smoothly.

For the archive *relay* problem, therefore, we believe it is more realistic to think about the situation from the point of view of the "archivist in the middle," who receives some digital content which has already been preserved for 50 years, and who must endeavor to ensure, even as the content is being used, that it is ready to be handed off for further preservation in another 50 years, if not sooner. The archive design question thus becomes:

> What kind of archive infrastructure can best facilitate the *ongoing* preservation of digital information, where a whole range of transitions at a variety of different levels will be required?

Archive Layers

In presenting the relay problem we have implicitly adopted as our fundamental approach to preservation the persistent archives strategy described by Thibodeau (2002), in which preserved information is maintained in hardware- and software-independent representations by loosely coupled, evolving components. Following Gladney (2007) and others, we conceptualize an archive as comprising three generalized architectural layers, each built atop the preceding one, with each layer being successively less generic and more application-specific than the preceding one:

A physical layer, that manages sequences of bits;

1. A logical layer, that manages objects and their interrelationships; and
2. An administrative layer, that implements an archive as a collection of objects, along with policies and end-user services.

Each of these layers is described in more detail below.

Physical Layer

The physical layer, or storage system, manages bit sequences independent of their interpretation. The minimum requirement for an archive physical layer is the ability to read and write an identified bit sequence, such as is provided by all computer file systems. At the physical layer, the archive problem reduces to guaranteeing the reliability of a storage system, and arranging for the bits to be copied to a new storage system when that guarantee can no longer be sustained. Note that any explicit or implicit (e.g., filename) semantics associated with the bit sequences are irrelevant at this layer, while the specific implementation of the storage system (online vs. offline, hierarchical vs. relational, etc.) is irrelevant to the higher layers.

Logical Layer

The logical layer, manifested as a repository system, interprets the archive's bit sequences as specific digital objects and as relationships between those objects. For example, a bit sequence maintained by the physical layer might be interpreted at the logical layer as a document in a specific format, where the relationship of the document to the format is maintained as well as the document itself. Note that the specific implementation of such a relationship is not specified—it could be represented in a filename (e.g., "document.pdf"), or as a "self-describing" format, or as a specific link between a document and a format specification. The key value added by the logical layer is the preservation of structure and semantics, whereas the physical layer preserves only the bits.

AdministrativeLlayer

The administrative layer groups collections of objects and their relationships into an archive; provides services (beyond the generic services provided by the underlying repository system) such as content-specific search and presentation; and enforces policies, such as selection, access, and maintenance. For example, a repository may contain video objects, but it is the responsibility of the administrative layer to provide a streaming video service for these objects, since such a service is sufficiently content-specific as to be unlikely to be provided by generic repository systems. Similarly, a storage system may announce a multi-year availability guarantee, but it is the responsibility of the administrative layer to monitor this guarantee and to take action when it fails or expires. The administrative layer's services and policies effectively define the archive.

Core Assumptions

Our fundamental assumption is that long-term digital preservation cannot be addressed by constructing individual archives that will last for a century or more. While individual archives can be long-lived, spanning multiple generations of storage media and software upgrades, a prudent conservator must assume that any archive will ultimately fail. If its contents are to be preserved, they must be migrated to other

archives, in possibly distant locations, using different technologies and under different administrative controls. Thus there is a need to define interfaces for the migration of content that can reach the greatest number of archives, including those not yet constructed or even conceived.

Curation Is Horizontal

In our layered architecture, the archiving problem becomes a scalability challenge, similar in many ways to the problem of designing a scalable communication network. In the case of archives, the necessary communication extends through long stretches of time, but the problem of interoperability is the same. As in networking, the solution lies in a stack of interfaces, assumed to be universal, which in the case of the archive stack means that they will be implemented by future as well as present systems.

An archive interface stack will support horizontal interoperability, whereby archives may migrate independently at the physical, logical, and administrative layers. For example, we can imagine a single archive being re-implemented atop a succession of physical layers; conversely, we can also imagine a single physical and logical implementation being repurposed by a succession of administrative interfaces and policies.

Resurrection Is MoreLikely than Immortality

Preservation must be cheap and easy. If we are to preserve digital information on a large scale, then the burden on providers must be small, preservation infrastructure must be flexible and adaptable, and multiple levels of preservation effort must be defined to accommodate varying archive resources. Preservation may be (and may often need to be) as minimal as crawling a website, saving the harvested files, and reflecting them back to the Web. This approach by itself does not guarantee that the archived information will remain usable by contemporary applications over time, but by capturing the contextual semantics of the harvested files, we can at least preserve the possibility of *resurrecting* full use of the information at any point in the future, assuming sufficient desire and resources.

Interfaces Should Be Minimal

Herring (2007) distinguishes between the "base and profile" and the "core and extension" metamodels for standards (and thus interface) definition. In the former case, all possible interactions are specified a priori, and then various profiles specify particular subsets to be implemented by different systems; whereas in the latter case, a minimal common core is specified, and then extended as needed. We believe archive interfaces must necessarily follow a "core and extension" metamodel. Archive migrations can then negotiate the set of extensions common to the source and destination layers, but can always fall back to the core functionality.

Archive Migration Today

Archive migration today is a human- and resource-intensive process characterized by the need to make content- and context-specific decisions and to perform one-time, ad hoc actions. In this section we look at the problems encountered at each architectural layer.

Physical Layer

Handing off bits from one archive to another, or from one storage system to another within an archive, is typically accomplished today by manually copying the bits. Storage area networks (SANs) are widely employed to virtualize underlying storage media and systems to good effect, making location, movement, and redundancy of data within such systems transparent to higher-level systems. However, SANs are typically proprietary and mutually incompatible, and in any case tend not to be deployed across institutional boundaries. The result is that content must copied, a process that, for large archives, can quickly test network capacity and bring up problems while dealing with large numbers of files (Shirky, 2005). In addition, to ensure that data are not silently corrupted by the copy operation, additional aggregation and packaging of the data, solely for the purpose of copying, are necessary (Boyko, Kunze, Littman & Madden, 2008; Library of Congress, 2008)

Logical Layer

Handing off archival objects from one repository system to another today requires repository- and content-specific schema matching.

The key component needed to support any type of object migration is a common, underlying object model to which each side can map. All repository systems are fundamentally derived from the Kahn-Wilensky Framework (Kahn & Wilensky, 1995), a good starting point, but the framework is too conceptual in nature to support direct interoperability.

The Metadata Encoding and Transmission Standard (METS) is an XML document format (and implied data model) for encoding aggregate digital library objects (McDonough, 2006). METS is certainly the right *kind* of specification to serve as an underlying object model considering its wide adoption and use as a common representation for submission, archival, and dissemination information packages within the OAIS archive reference model (Consultative Committee for Space Data Systems [CCSDS], 2002). However, METS is also a very flexible and general standard, a characteristic that has given rise to the definition of many application profiles of METS[1] and the inevitable, concomitant decrease in interoperability (McDonough, 2006). METS-producing applications find it easy to map their internal data models to METS, but METS-consuming applications are typically restricted to accepting only the profile(s) within their application domain.

This is particularly true in the case of repository systems. The well-known repository systems in use today, Fedora and DSpace — Greenstone, too, though not traditionally thought of as a repository — define and support METS profiles that closely correspond to their internal data models. Thus, for example, the Fedora METS profile (Fedora, 2007a) is essentially a syntactic re-encoding of Fedora's internal data model, FOXML (Fedora, 2007b), and analogously for DSpace's METS profile (Wolfe & Reilly, 2008). Examining the differences between these repositories' respective profiles, that is to say, the differences between their internal data models, it is clear that some differences are purely syntactic—that is, a constant transformation can be applied that is independent of the content being mapped. But there are other differences that are content- and context-dependent, and that therefore require human

[1] METS Profiles http://www.loc.gov/standards/mets/mets-profiles.html

judgment for amelioration. The difficulty of this task has been documented in real-world experiences in cross-repository mapping (Emly, 2007).

We conclude that handing off objects across repository systems today is certainly possible, but is not sufficiently automated or automatable.

Administrative Layer

Handing off whole archives today can be likened to a treaty negotiation in which specialists from each side (curators, content and metadata specialists, system architects, programmers, system administrators) meet and negotiate the terms and mechanisms of the transfer. Rank and McDonald (2005) report on the difficulties in handing off a large, operational archive of remote-sensing imagery operated by one U.S. government agency to an archive operated by another. They also note the limited help provided by the OAIS archive reference model (CCSDS, 2002) in performing such a handoff.

Relay-supporting Archive Migration

What would archive migration look like with interoperability mechanisms in place that supported easy migration, and by extension, easy preservation relays over extended periods of time? In this section we again examine migration at the three architectural layers, but now consider requirements for archive handoff interoperability as well as existing efforts to implement such interoperability.

Physical Layer

To support handoffs of bits across storage systems, across repository systems, and across institutions, the storage infrastructure implementing the physical layer must provide the same kinds of functionality now provided by SANs—virtualization of storage, replication, and policy-based control—on an unprecedentedly large (ideally global) scale. Because this implies crossing the boundaries of storage systems that are typically proprietary, and enabling institutions to share resources and interoperate in new ways, the storage infrastructure must be defined by open protocols.

Lots of Copies Keep Stuff Safe (LOCKSS) (Rosenthal & Reich, 2000) is one protocol-based approach to storage. It allows cooperating institutions to replicate files across distributed, dedicated storage depots; additional protocols monitor for corruption. Although initially targeted at electronic journal publishers and libraries, LOCKSS is increasingly being used for more general file replication within private networks.

Another technology, Logistical Networking (LN) (Beck, Moore & Plank, 2002), is the most explicit attempt to date to apply the Internet's architectural approach to storage. The key to the Internet's design is an "hourglass" architecture, at the narrow waist of which is a highly generic, common service—the IP protocol for best-effort datagram delivery—that mediates between basic shared physical resources (network bandwidth in the Internet's case) and the applications that want to use those resources. Protocols built on top of IP, such as TCP, provide higher-level functionality such as reliable communication and persistent connections. LN's basic elements closely track this design. At the narrow waist of LN is the Internet Backplane Protocol (IBP), which

mediates (only) best-effort, relatively short-term storage leases. Higher-level protocols built on IBP provide persistent and replicated storage, abstractions such as files and filesystems, and so forth.[2]

Logical Layer

A common, underlying object model is needed to support object migration, a model that allows objects to be handed off from repository system to repository system unrestricted by application domain profiles or the need for human judgment. To satisfy this goal, we argue that it is both necessary and sufficient for the object model to represent the following entities:

1. Bitstreams, to hold object content;
2. Objects, as an aggregative mechanism for bitstreams;
3. Persistent, universally unique identifiers for objects, for naming; locally unique identifiers for bitstreams, for disambiguation;
4. Fixity metadata for bitstreams and objects, to support end-to-end reliability;
5. Object and bitstream semantics, and persistent associations to those semantics, to support resurrectability; and
6. Inter-object relationships, to model ontological assertions.

The sufficiency of this list derives from the modesty of our goal: handing off an object from one repository to another. The ability of an archive to deal sensibly with an object will hinge on its ability to recognize and interpret the object's bitstreams and formats; but in an extended relay across time, there is no requirement that *every* archive understand the object, only that it is possible for *some* archive in the future to do so.

Notably missing from the above list is any mention of metadata beyond fixity information. The AIHT experiment concluded that metadata requirements are really desires, or as Shirky (2005) puts it, "requirements aren't." Moreover, from our own experience, mapping descriptive metadata is not entirely automatable, but must be customized on at least a collection-by-collection basis due to wide variability in metadata quality and interpretation (Janée & Frew, 2005).

The Pathways Core data model (Warner et al, 2007), precursor to the Object Reuse and Exchange (ORE) data model (Van de Sompel & Lagoze, 2007) currently under development, was the first effort at creating a cross-repository object model. It nicely satisfies most of the above requirements, with a caveat noted below. Gladney (2004) presents a similar data model that incorporates cryptographic signing to support trust assertions.

Pathways Core relies on external format registries for semantic definitions: a bitstream's interpretation is defined by reference to a (single) format in a registry. This approach, while admittedly ubiquitous, has two limitations from our perspective of long-term preservation. First, as we explore in more detail in the appendix, there are some types of information, notably scientific datasets, that require a web of inter-related objects for their interpretation and use. Format registries may be sufficient for textual, video, and audio documents, where a document's format specification provides sufficient information to resurrect a visual or audible *rendering* of the

[2] We are currently developing a prototype archive built on an LN-based distributed storage network.

document, but remote-sensing imagery requires far more contextual and provenance information to support its use in scientific modeling. Second, Pathways Core bitstream semantics are represented as references, but what is at the end of those references is undefined and outside the data model's scope. The implicit assumption is that the format registry will exist forever. In reality, though, the registry is an archive like any other, a participant in its own relay.

The National Geospatial Digital Archive (NGDA) data model (Janée, Mathena & Frew, 2008) attempts to address these shortcomings. In the NGDA data model, the interpretation of both objects and bitstreams is defined by one or more "definition" relationships, but here the relationships must explicitly target other archival objects. Thus the NGDA data model replaces the bifurcated view of archives of objects on the one hand, referencing registries of formats on the other, with an undifferentiated sea of inter-related archival objects residing across a federation of archives. Application and testing of this data model is ongoing.

Despite the lack of a proven, universal object data model today, we believe one is within reach.

Administrative Layer

What is needed to support easy whole-archive migration above and beyond the aforementioned interoperability mechanisms at the physical and logical layers? Actually, very little. Any source archive will implement policies related to content selection, ingest, management, and so forth, but while such policies may be of documentary interest to a receiving archive, their migration is not necessary; any receiving archive will implement its own policies anyway. Any source archive will also provide services such as discovery and content-specific access, but again, the receiving institution will need and want to provide its own services anyway. We concur with the AIHT experiment's advocacy of a data-centric approach to migration: "A data-centric strategy assumes that the interaction between institutions will mainly be in the passing of a bundle of data from one place to another—that data will leave its original context and be interpreted in the new context of the receiving institution." (Shirky, 2005)

Given a strict focus on archived content, then, the only information needing handoff at the administrative layer is a root object or starting crawl point, analogous to the super block on a disk drive. This idea has been explored in work on using the OAI-PMH protocol to automatically harvest the content of OAIS-compliant repositories (Bekaert & Van de Sompel, 2005).

To a starting crawl point we would add a whole-archive dependency descriptor that describes an archive's external dependencies, in particular, its dependencies on any other archives (including format registries) and on any other systems (including persistent identifier resolution systems). This would enable the receiving institution to see at a glance that a source archive is dependent on, for example, the PRONOM format registry for semantic definitions and the DOI system for identifier resolution.

Conclusions

Existing technologies come close to, but fall short of, implementing the kinds of interoperability needed to support easy migration of preserved content across storage systems, across repository systems, and across curators and institutions. At the physical layer, Logistical Networking, if adopted as widely as other Internet protocols, could change how we conceive of and use storage. It would take functionality that is currently commonly available on local scales only—bit movement and replication automated to the extent that storage actions are expressible as simple, declarative policy and ownership changes—to a global scale. At the logical layer, a standard, uniform data model for information and semantics would remove the ingest barriers that currently exist between repository systems. And at the administrative layer, standardized crawling points and whole-archive dependency descriptors would facilitate institutional turnovers while mitigating the risk that preserved content gets "dropped" at critical transition times.

The net result of relay-supporting interoperability is a baseline level of digital preservation. The myriad challenges posed by curation of digital content will remain with us, of course—all the problems of selection and identification, of format obsolescence over time, of providing search, access, and other services, and so on—but at least the risk of fundamental information loss is reduced.

Proposed interoperability solutions must be tested. At the physical layer, it must be possible to implement repository systems on top of a common, protocol-defined storage substrate, and it must be possible to move and replicate data across repository and institutional boundaries. Furthermore, institutions must have faith in the storage substrate's promises of reliability, ownership, and privacy. Specific tests might include replicating data across dissimilar storage systems within an archive; migrating a source archive to a destination archive without transferring any content bits; and changing ownership and replication characteristics through policy changes.

The logical layer would require the most testing, as defining universal data models is notoriously difficult. The data model must be implemented by several repository systems, and handoffs must be tested for a variety of object types and subject domains. A specific, challenging test would be to migrate a type of content which the destination archive was never intended to accommodate: for example, migrating a collection of video content to a geospatial data archive. The destination repository may not be able to provide any meaningful services over the newfound content other than raw access, but if it is capable of handing off the content in turn to a third archive that *can* provide these services, then the relay principle is proven.

At the administrative layer, migration testing must mimic institutional handoffs, particularly in that most pessimistic scenario, when the source institution is unable to provide any guidance or give any support to the receiving institution.

Essentially, we are suggesting that it is time for a second Archive Ingest and Handling Test (Shirky, 2005). But whereas the first test left issues of technology open, a second test would focus on a specific set of candidate interoperability technologies.

Appendix

In our work on long-term preservation of geospatial data we have found that certain types of geospatial datasets present challenging problems to preservation beyond their large size.

First, the contextual information required to use the datasets can be quite complex. For example, using remote-sensing imagery in scientific models requires detailed knowledge of platform and sensor characteristics, and calibration and processing algorithms. Strictly speaking, such contextual information constitutes metadata, but in practice, being too large and complex, it is not handled as such. Instead of being bundled in a metadata record, the information is typically held in external documents and websites. Of course, from a preservation perspective, the contextual information must be co-archived with the data regardless of its source.

Second, certain types of geospatial datasets, and here again we will focus on remote-sensing imagery, require periodic reprocessing. For example, the Sea-viewing Wide Field-of-view Sensor (SeaWiFS) dataset has already been reprocessed eight times in ten years.[3] Many of these reprocessings corrected instrumentation problems that were discovered only in the course of using the data, implying that more such reprocessings may be required in the future. But even if we assume that the basic instrumentation will be considered "fixed" at some point in the future, reprocessings will still be needed due to improved scientific models and understanding, such as an improved model of solar irradiance in the case of one SeaWiFS reprocessing. Periodic reprocessing is a necessity for any type of climate data record (Colton, Karl, Goldberg & Bates, 2003).

Furthermore, reprocessing a remote-sensing dataset requires information beyond even the contextual information described above. In addition to the "raw" data that served as the original source for the product in need of reprocessing, reprocessing can require scientific papers, algorithm documentation, processing source code, calibration tables and databases, and even ancillary datasets (Linda, 2006).

Putting these problems together—complex contextual information, multiple datasets related by provenance chains and workflows, and the need to periodically modify and exercise those workflows—we conclude that an archive data model must represent information not as a set of independent objects, but as a graph of interdependent objects linked by dependency, provenance, and other types of relationships.

We have encountered these preservation challenges with geospatial data, but we believe similar challenges will occur with any type of scientific data for which provenance of the data represents an important aspect of its use.

[3] Ocean Color Data Reprocessing http://oceancolor.gsfc.nasa.gov/REPROCESSING/

References

Beck, M., Moore, T., & Plank, J.S. (2002). An end-to-end approach to globally scalable network storage. *ACM SIGCOMM Computer Communication Review 32* (4*)* pp. 339–346. doi:10.1145/964725.633058

Bekaert, J., & Van de Sompel, H. (2005). Access interfaces for open archival information systems based on the OAI-PMH and the OpenURL Framework for context-sensitive services. *Ensuring Long-term Preservation and Adding Value to Scientific and Technical data (PV 2005).* Edinburgh, UK, November 21–23, 2005. Retrieved November 17, 2008, from http://www.ukoln.ac.uk/events/pv-2005/pv-2005-final-papers/032.pdf

Boyko, A., Kunze, J.A., Littman, J., & Madden, L. (2008). *The BagIt File Package Format*. Version 0.95 (July 11, 2008). Retrieved November 17, 2008, from http://www.cdlib.org/inside/diglib/bagit/bagitspec.html

Colton, M., Karl, T.R., Goldberg, M.D., & Bates J.J. (2003). *Creating climate data records from NOAA operational satellites*. National Oceanic and Atmospheric Administration (NOAA) white paper. August 2003. Retrieved November 17, 2008, from http://cimss.ssec.wisc.edu/itwg/groups/climate/Creating_CDRs_from_NOAA_Satellites_White_Paper_18_Aug.pdf

Consultative Committee for Space Data Systems (2002). *Reference model for an Open Archival Information System (OAIS)*. CCSDS 650.0-B-1, Blue Book (January 2002). Retrieved November 17, 2008, from http://public.ccsds.org/publications/archive/650x0b1.pdf

Emly, M. (2007). *MIDESS project final report*. Retrieved November 17, 2008, from http://ludos.leeds.ac.uk/midess/MIDESS-final-report.pdf

Fedora Project (2007a). *Encoding fedora objects in METS*. Fedora Release 2.2.1, August 1, 2007. Retrieved November 17, 2008, from http://www.fedora.info/download/2.2.1/userdocs/digitalobjects/rulesForMETS.html

Fedora Project (2007b). *Introduction to Fedora Object XML (FOXML)*. Fedora Release 2.2.1, August 1, 2007. Retrieved November 17, 2008, from http://www.fedora.info/download/2.2.1/userdocs/digitalobjects/introFOXML.html

Gladney, H.M. (2004). Trustworthy 100-year digital objects: Evidence after every witness is dead. *ACM Transactions on Information Systems (TOIS) 22*(3), pp. 406–436. July, 2004. Retrieved November 17, 2008, from doi:10.1145/1010614.1010617

Gladney, H.M. (2007). *Preserving digital information*. Berlin; New York: Springer Verlag. ISBN 3-540-37886-3

Hedstrom, M. (2001). *Exploring the concept of temporal interoperability as a framework for digital preservation*. Third DELOS Workshop on Interoperability and Mediation in Heterogeneous Digital Libraries (Darmstadt, Germany; September 8–9, 2001). Retrieved November 17, 2008, from http://www.ercim.org/publication/ws-proceedings/DelNoe03/10.pdf

Hedstrom, M., et al. (2003). It's about time: Research challenges in digital archiving and long-term preservation. *Final Report, NSF Workshop on Research Challenges in Digital Archiving: Towards a National Infrastructure for Long-term Preservation of Digital Information* (Warrenton, VA; April 12–13, 2002). Retrieved November 17, 2008, from http://www.si.umich.edu/digarch/NSF%200915031.pdf

Herring, J.R. (2007). *Annex to OGC policies and procedures – The specification model – Structuring an OGC specification to encourage implementation*. OGC discussion paper OGC 07-056r1, version 0.0.9. Open Geospatial Consortium, Inc. (July 12, 2007). Retrieved November 17, 2008, from http://portal.opengeospatial.org/files/?artifact_id=21976

Janée, G., & Frew, J. (2005). A hybrid declarative/procedural metadata mapping language based on python. *Research and Advanced Technology for Digital Libraries: Proceedings of the 9th European Conference (ECDL)*. (Vienna, Austria; September 18–23, 2005): pp. 302–313. Retrieved November 17, 2008, from doi:10.1007/11551362_27

Janée, G., Mathena, J., & Frew, J. (2008). A data model and architecture for long-term preservation. *Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL)* (Pittsburgh, PA; June 16–20, 2008):134–144. Retrieved November 17, 2008, from doi:10.1145/1378889.1378912

Kahn, R., & Wilensky, R. (1995). A framework for distributed digital object services. Republished in *International Journal on Digital Libraries, 6*(2) (April 2006): pp. 115–123. Retrieved November 17, 2008, from doi:10.1007/s00799-005-0128-x

Library of Congress. (2008, June). *Digital preservation newsletter*. Retrieved November 17, 2008, from http://www.digitalpreservation.gov/news/newsletter/200806.pdf

Linda, M. (2006). OMPS aggregation and packaging. *2006 CLASS Users' Workshop* (Boulder, CO; August 7–8, 2006). Retrieved November 17, 2008, from http://ngdc.noaa.gov/dmsp/2nd_class_workshop/class.html

McDonough, J.P. (2006). METS: Standardized encoding for digital library objects. *International Journal on Digital Libraries 6*(2) (April 2006): pp.148–158. Retrieved November 17, 2008, from doi:10.1007/s00799-005-0132-1

Rank, R. & McDonald, K.R. (2005). A NOAA/NASA pilot project for the preservation of modis data from the earth observing system (EOS). *Ensuring Long-term Preservation and Adding Value to Scientific and Technical data (PV 2005)* (Edinburgh, Scotland; November 21–23, 2005). Retrieved November 17, 2008, from http://www.ukoln.ac.uk/events/pv-2005/pv-2005-final-papers/037.pdf

Rosenthal, D.S.H., & Reich, V. (2000). Permanent web publishing. *Proceedings of the FREENIX Track: 2000 USENIX Annual Technical Conference* (San Diego, CA; June 18–23, 2000):129–140. Retrieved November 17, 2008, from http://www.usenix.org/events/usenix2000/freenix/full_papers/rosenthal/rosenthal.pdf

Rosenthal, D.S.H., Robertson, T., Lipkis, T., Reich, V., & Morabito, S. (2005, November). Requirements for digital preservation systems: A bottoms-up approach. *D-Lib Magazine 11*(11). Retrieved November 17, 2008, from doi:10.1045/november2005-rosenthal

Ross, S., & Hedstrom, M. (2005). Preservation research and sustainable digital libraries. *International Journal on Digital Libraries 5*(4) (August 2005): pp. 317–324. Retrieved November 17, 2008, from doi:10.1007/s00799-004-0099-3

Shirky, C. (2005). *Library of Congress archive ingest and handling test (AIHT) final report*. Retrieved November 17, 2008, from http://www.digitalpreservation.gov/library/pdf/ndiipp_aiht_final_report.pdf

Thibodeau, K. (2002). Overview of technological approaches to digital preservation and challenges in coming years. *The State of Digital Preservation: An International Perspective* (Washington, DC; April 24–25, 2002). Retrieved November 17, 2008, from http://www.clir.org/pubs/reports/pub107/thibodeau.html

Van de Sompel, H., & Lagoze, C. (2007). Interoperability for the discovery, use, and re-use of units of scholarly communication. *CTWatch Quarterly 3*(3) (August 2007): pp. 32–41. Retrieved November 17, 2008, from http://www.ctwatch.org/quarterly/articles/2007/08/interoperability-for-the-discovery-use-and-re-use-of-units-of-scholarly-communication/

Warner, S., Bekaert, J., Lagoze, C., Liu, X., Payette, S., & Van de Sompel, H. (2007). Pathways: Augmenting interoperability across scholarly repositories. *International Journal on Digital Libraries 7*(1/2) (October 2007): pp. 35–52. Retrieved November 17, 2008, from doi:10.1007/s00799-007-0016-7

Wolfe, R., & Reilly, W. (2008). *DSpace METS document profile for submission information packages (SIP)*. Dated May 2, 2008. Retrieved November 17, 2008, from http://wiki.dspace.org/index.php/DSpaceMETSSIPProfile