

Self-Assessment of a Long-Term Archive for Interdisciplinary Scientific Data as a Trustworthy Digital Repository

Robert R. Downs and Robert S. Chen

Center for International Earth Science Information Network (CIESIN)

The Earth Institute, Columbia University

61 Route 9W, Palisades, NY 10964

{rdowns, bchen} @ciesin.columbia.edu

Abstract

Long-term preservation and stewardship of scientific data and research-related information are vitally important to future science and scholarship. Scientific data archives can offer capabilities for managing and preserving disciplinary and interdisciplinary data for research, education, and decision-making activities of future communities of users. Meeting the requirements for a trusted digital repository will help to ensure that today's collections of scientific data will be available in the future. A continuing self-assessment of a long-term archive for interdisciplinary scientific data is being conducted to identify the additional steps needed for it to become a trustworthy repository. Recommendations include a strategy for collaborative organizational sustainability, a model for submission and workflow to ingest interdisciplinary scientific data into a repository, and a plan for facilitating intra-organizational transfer between repositories.

Keywords

scientific data stewardship, trustworthy digital repository, long-term archive, digital preservation, data curation

1. New Challenges for Scientific Data Stewardship

Today, scientific data are routinely created in digital form and analyzed using computer-based applications. In addition to enabling the creation and analysis of scientific data, the digital form facilitates much greater data sharing and reuse by others, which has led to the spread of data-driven science practices within scientific communities (Lesk, 2008). Rapid expansion in the collection and use of digital data is one major element in the more general development of eScience and eSocial science (Arms, Calimlim, & Walle, 2009). Wider access to scientific data over the Internet has also enabled the development of

digital resources and web-based services that facilitate uses of data beyond those that may have been envisioned by the original data producers. For example, more widespread access to scientific data and analytical tools presents new opportunities for creating learning objects in support of both formal and informal education (National Science Foundation Cyberinfrastructure Council, 2007).

Along with these new opportunities for sharing and using scientific data come new challenges for scientific data stewardship. Some scientific practices are evolving rapidly to leverage the opportunities offered by digital data, but other practices have not kept up. For example, it is now much easier and more common for scientists to develop and analyze primary data, and generate secondary datasets and publication-quality visualizations of their data, entirely on their personal computers, without using any shared computing resources maintained by a computer professional or information specialist (except perhaps an Internet connection). As a result, practices such as more careful labeling, organization, and documentation of data files and regular backups have become less prevalent and more haphazard (Cocker, 2005; Marshall, Bly, and Brun-Cottan, 2006). Moreover, in many fields of research, scientists are tapping a wider variety of primary and secondary databases in their work, and may themselves generate a much larger number and variety of datasets and revised versions of datasets in their careers. Yet practices and procedures for citation of data, application of unique data identifiers, and version control are still under development and not yet widely used in most scientific disciplines.

Challenges of this type are also evident at the organizational level. On the one hand, scientific data centers, libraries, government agencies, and other groups have moved rapidly to online digital data access and services, drastically reducing usage of traditional offline access methods. On the other hand, practices for storage and preservation of these digital data resources are far from the maturity and reliability achieved for traditional non-digital media. Technical challenges include the limited shelf-life of storage media and the rapid evolution of computer technologies. However, equally if not more important are the institutional and organizational challenges of dealing with digital data in the long term, such as the uncertain longevity of the relatively new organizations that now produce and archive many key datasets, the increasing complexity of intellectual property issues associated with these data, and the proliferation of different digital data standards and formats that in many instances require specialized knowledge and tools.

In an effort to meet these challenges and improve their capabilities for stewardship of digital resources, many organizations are now beginning to implement preservation environments, such as digital repositories, to manage their collections of digital information, including scientific data. Such preservation environments offer a way for

organizations to manage the authenticity and integrity of their digital resources over time (Moore, 2008).

Similar to the intellectual property assets and electronic records managed by institutional archives and repositories, the contents of scientific archives and repositories need to be trustworthy (Gladney, 2006). If the scientific data and research-related information stored in a digital repository cannot be trusted, then their value for future use becomes questionable. In the case of scientific data, it is essential that trust encompasses not only the integrity of the digital data, but also the authenticity of the links between the data and the data sources and documentation.

The Open Archival Information Systems Framework (CCSDS, 2002) offers guidance for the long-term stewardship of scientific data and other digital resources. Based on this standard, instruments have been developed for assessing the trustworthy nature of data archives and other digital repositories. These instruments can assist managers to conduct self-assessments to identify weaknesses in and help improve capabilities for long-term stewardship of digital collections. One of the tools developed for this purpose is the Trusted Repositories Audit & Certification: Criteria and Checklist (TRAC) document (OCLC and CRL 2007). The TRAC describes a set of criteria for a trustworthy repository in three categories, “Organizational Infrastructure”, “Digital Object Management”, and “Technologies, Technical Infrastructure, & Security”. We are using the TRAC to conduct a continuing self-assessment of a long-term archive that was established for the preservation of interdisciplinary scientific data.

2. Conducting the Self-Assessment of a Long-Term Archive

The NASA Socioeconomic Data and Applications Center (SEDAC), operated by the Center for International Earth Science Information Network (CIESIN) of Columbia University, produces, archives, and disseminates scientific data and offers services to improve understanding of human interactions in the environment. A variety of user communities rely on SEDAC to continually provide access to scientific data and services in support of research, education, and decision making (Downs and Chen, 2003). SEDAC has been characterized as a “reference collection” serving “large large segments of the general scientific and education community” by the National Science Board (2005, Appendix D).

As an operational archive within NASA’s Earth Observing System Data and Information System (EOSDIS), SEDAC does not have an explicit responsibility for long-term archiving. CIESIN has therefore chosen to develop and implement a SEDAC Long-Term Archive (LTA) in collaboration with the Columbia University Libraries and University’s

Earth Institute, of which CIESIN is a unit (Downs, Chen, Lenhardt, Bourne, & Millman, 2006). The LTA is being established initially as a distinct archive within CIESIN parallel to the active SEDAC digital data archive. In the event that SEDAC were to cease operations, the LTA would serve as a long-term home for important and unique SEDAC data holdings. If CIESIN or the Earth Institute were to cease operations or lack the resources to maintain the archive, the Columbia Libraries have agreed to assume responsibility for the LTA as part of its own long-term digital repository, currently under development. An LTA Board with members from CIESIN, the Earth Institute, and the Columbia Libraries oversees the implementation and operation of the SEDAC LTA.

Like other organizations that have evaluated alternative digital repository platforms (Groenewegen & Treloar, 2008; Marill & Luczak, 2009), CIESIN has selected the open source software suite Fedora, the Flexible Extensible Digital Object Repository Architecture (Lagoze, Payette, Shin, & Wilper, 2006), after assessing alternative systems and testing Fedora in a pilot project. CIESIN is using Fedora as the basis for the LTA's digital repository and asset management system, in conjunction with the VITAL software from VTLS, Inc.

The LTA Board recommended a self-assessment of the LTA as an essential step for the LTA to meet the requirements for certification as a trusted digital repository, to identify areas in which the management and operation of the LTA could be improved, and to guide future development of the LTA, based on emerging standards. The TRAC document was chosen as the initial instrument for the self-assessment for several reasons. First, the TRAC is based on the OAIS framework and contains a broad set of evaluation criteria for assessing a digital repository. As described by Ambacher (2007), the TRAC was developed in an open manner, in consultation with the community most concerned with digital preservation, including the Research Libraries Group (RLG) and the U.S. National Archives and Records Administration (NARA). In addition, the TRAC is one of the primary resources being used by the Digital Repository Audit and Certification Working Group of the Consultative Committee for Space Data Systems to establish an international standard of metrics for digital repository audit and certification.

We are using the TRAC to conduct the self-assessment on a continuing basis to identify areas for further improvement and to check on past changes in processes and procedures. As the LTA continues to evolve and adopt new technologies and practices, relevant sections of the TRAC document will be revisited. Because an international standard does not yet exist, we may consider using elements of other audit tools in the future to supplement the TRAC, especially as digital repository practices mature and requirements for the curation, stewardship, and preservation of digital resources evolve in response to new needs and approaches.

The self-assessment has been conducted by reviewing each TRAC requirement and analyzing the relevant LTA policies, plans, and procedures to determine whether the requirement has been met, in whole or in part. In some cases, small adjustments are made immediately in these policies, plans, and procedures. Major changes are subject to approval by the LTA Board, and all changes are carefully documented. In addition to analysis of individual requirements, the current technological infrastructure, organizational capabilities, and relevant documentation also are being reviewed in terms of the overall set of TRAC requirements. Conducting the self-assessment in this manner has provided the opportunity to identify areas in which other archives and repositories might consider making improvements.

3. Initial Results of the Self-Assessment

The initial self-assessment has found that the SEDAC LTA meets the TRAC criteria with respect to traditional scientific data management practices and implemented digital repository capabilities. These results are summarized in Table I.

	Policies	Plans	Procedures	Forms	Documentation	Contracts
Relevant	4	6	4	2	2	1
Improved	1	2	3	0	0	0

Table 1. Relevant Resources Identified and Improved During Initial Self-Assessment

Several types of written resources were analyzed during the initial self-assessment of the SEDAC LTA to determine their relevance to the TRAC criteria. The resources that were analyzed included drafts as well as documents that had been formally approved. Resources relevant to meeting a specific criterion were listed as evidence for that item. Nineteen resources that were identified as relevant to meeting specific TRAC criteria were listed as evidence for meeting the items to which they pertained. The categories of resources identified as relevant to the self-assessment of the SEDAC LTA include policies, plans, procedures, forms, documentation, and contracts.

The self-assessment also provided an opportunity to revise some of the relevant resources to meet the TRAC criteria and to improve the SEDAC LTA. Six resources were improved during the self-assessment, within the categories of policies, plans, and

procedures. Each of the improved resources was in a draft state and under review, which the self-assessment fostered by enabling the identification of specific areas needing further improvement.

The self-assessment has also revealed three areas where the SEDAC LTA has made progress but needs additional development. The three areas are:

1. a strategy for collaborative organizational sustainability;
2. a model for submission and workflow to ingest interdisciplinary scientific data into a repository; and
3. a plan for facilitating intra-organizational transfer between the repository at CIESIN and the repository managed by the Libraries.

These are also areas where the SEDAC LTA experience may be instructive for other archives, repositories, and scientific data centers.

3.1 Strategy for Collaborative Organizational Sustainability

Many digital data archives like SEDAC have emerged in recent decades that do not have an explicit mission to preserve digital data and information in the long run. Data are developed, managed, and disseminated primarily to meet ongoing research and operational needs. Even if the utility of a data collection or database increases over time as more data are acquired and integrated, funding is rarely guaranteed to maintain and preserve the data in perpetuity. Moreover, even when funding organizations appear to have made long-term commitments to data support and stewardship, such commitments may only last if the organization itself persists. Unfortunately, even government agencies in stable governments sometimes cease to exist, or their missions or funding are significantly changed. Similarly, foundations, universities, libraries, museums, and other private sector organizations are not immune to economic and institutional upheaval. Figure 1 illustrates the different lifetimes of some major U.S. universities, government agencies, and other nongovernmental organizations. A number of private U.S. universities have had more than 250 years of continual operation as centers of knowledge preservation and dissemination.

Nevertheless, recognizing that no organization can absolutely guarantee long-term preservation and access does not mean that reasonable strategies for organizational sustainability are not feasible. Development of collaborations within and between institutions and associated contingency plans provides viable options for the long-term survivability of data and continued access. Such collaborative partnerships can reduce the dependence on limited resources and enable common approaches to meet the goals of collaborating partners.

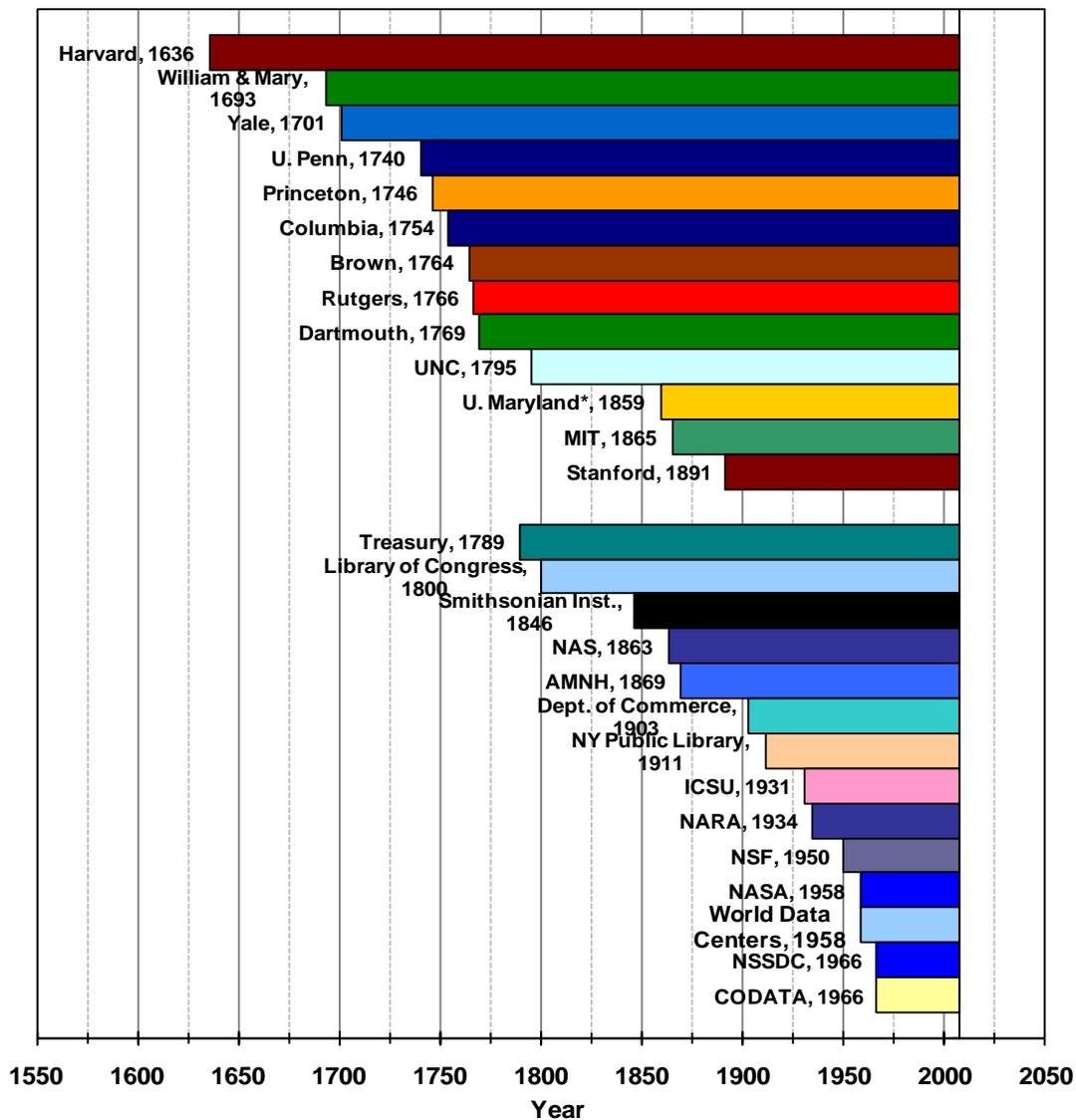


Figure 1. Longevity of Selected Universities, Government Agencies, and Other Institutions by Year Established.*

* Acronyms: MIT = Massachusetts Institute of Technology; NAS = National Academy of Sciences; AMNH = American Museum of Natural History; ICSU = International Council for Science; NARA = National Archives and Records Administration; NSF = National Science Foundation; NASA = National Aeronautics and Space Administration; NSSDC = National Space Science Data Center; CODATA = Committee on Data for Science and Technology

Collaboratively developing and managing the SEDAC LTA with the Columbia University Libraries and the Earth Institute provides the basis for what we believe will be a sustainable organizational infrastructure for the archive, including both its technical infrastructure and the content of its collection. Columbia University has a long-term mission to “advance knowledge and learning at the highest level and to convey the products of its efforts to the world.” (Columbia Mission Statement, http://www.columbia.edu/about_columbia/mission.html). The Columbia Libraries have identified the need to “increasingly embrace the challenges of ensuring the long-term availability of digital resources” and have specifically singled out collaboration with CIESIN and SEDAC as an important step in the development of Columbia’s Long-Term Digital Archiving Service (Columbia University Libraries, 2006, pp. 12-13). In the long run, it is clear that the SEDAC LTA holdings will be a tiny portion of a much larger collection of digital assets held and preserved by the University, but until that time it can serve as an important test case of how a reference collection of scientific data can be integrated into a large digital archive.

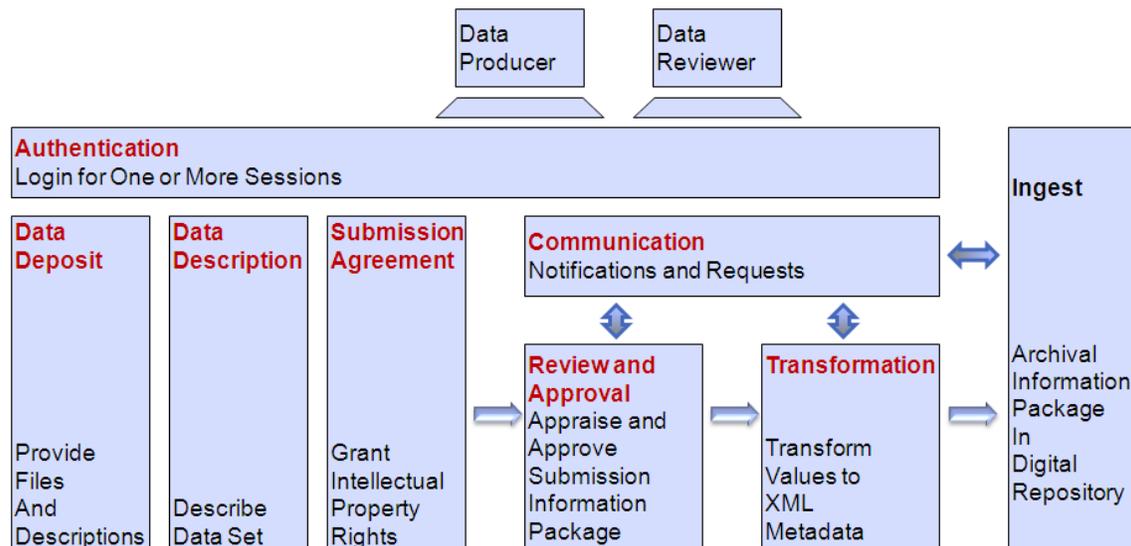
As noted previously, the SEDAC LTA Board includes representatives from SEDAC, the Columbia University Libraries, and the Earth Institute. The Board determines the appraisal criteria for accession and decides whether nominated scientific data sets will be accessioned. In making these decisions, the Board is cognizant of the potential long-term costs of digital preservation. Currently, the SEDAC LTA is managed by SEDAC staff. In the event that sponsorship discontinues for the operation of the SEDAC, contingency plans change the composition of the Board and the management of the archive so that the designated organizational entities can assume control and accept the responsibility for continued management and operation as needed. This will be feasible if the costs of transitioning the LTA holdings to Columbia’s larger digital collection are small. By developing the current archive in accordance with accepted standards and using flexible, open source tools, we believe that we will be able to ensure that the transition costs are kept to a minimum.

3.2 Model of Submission and Workflow for Preserving Scientific Data

A key challenge for most data archives is gathering the information needed to meet preservation metadata needs. When data submission and metadata creation are separated in time and carried out by different individuals, ensuring accuracy and consistency can be extremely difficult as well as costly. Therefore, we believe that it is essential to develop mechanisms whereby suitable data descriptions and metadata are captured when data resources are actually submitted to the repository, to the extent possible. Improving capabilities for producers to submit scientific data and associated metadata to a repository soon after creation should not only increase the quality of the metadata available, but also improve the efficiency of the process. For example, review and appraisal of submitted

resources can be incorporated into the overall workflow, helping to reduce duplication of effort during the pre-ingest process. A systematic approach also can ensure that validation, both manual and automated, is completed effectively and efficiently (Consultative Committee for Space Data Systems, 2004).

A model has been developed to guide the design of capabilities for web-based submission and workflow for ingest of interdisciplinary scientific data to the repository. The model specifies functional capabilities to support data submission services and addresses the TRAC criteria for submission, review, preparation, and ingest. We have begun customizing the open source submission software, VALET, to meet these specifications. Identification of successful practices will be used to inform the design of the data submission and workflow system and user testing will be conducted to identify additional enhancements that are needed. The model is presented in Figure 2.



Adapted from: Downs and Chen. 2008. Creating a Trustworthy Digital Repository for a Long-Term Archive of Interdisciplinary Data: A Case Study. 21st International CODATA Conference, 5-8 October, 2008 Kyiv, Ukraine.

Figure 2. Model of Submission and Workflow for Preserving Scientific Data

3.3 Plan for Intra-Organizational Collection Transfer

Establishing capabilities for transferring collections between repositories is an important way to mitigate the risks associated with any specific host repository. Planning for transfer capabilities enables a repository to address vulnerabilities associated with its current location and organizational framework. The ability to transfer objects between repositories distributes the risk associated with the transferring repository infrastructure

or its future capabilities (Caplan, 2008; Janée, Frew, & Moore, 2009; Littman, 2009). Reducing the costs of transfers through automation and use of standards is necessary to ensure that transfers are actually carried out even when a repository is closed under stressful or resource-restricted conditions. It is also important for transfers to occur with high reliability, since the opportunity to go back and fix problems may be limited once a transfer takes place.

The Columbia University Libraries are currently implementing Fedora for its long-term digital repository operations. The Libraries and the SEDAC LTA have recognized that advance planning and testing are needed to ensure the maximum possible interoperability and reliability between their systems. We are therefore developing plans for a series of tests of the two-way transfer of selected scientific data sets between the two repositories. The results of these tests will assist in the evaluation of the requirements to facilitate future integration of the repository environments.

4. Conclusions

The self-assessment of the SEDAC LTA has identified several important challenges and possible strategies for scientific data archives and repositories. Archives and repositories could benefit from continued self-assessment to ensure that they meet established criteria for trustworthiness, especially during the transition of infrastructure and collections to digital repository systems. Continuous assessment and improvements are needed to ensure that the trustworthiness of data and metadata are maintained as collections grow, as new technologies are adopted, and as new services are offered for current and future user communities. It is likely that such assessments would also be an important step towards certification of digital data archives and repositories for trustworthiness, should formal certification standards and criteria be established by the relevant communities in the future.

5. Acknowledgements

This article is based on the presentation, Conducting a Self-Assessment of a Long-Term Archive for Interdisciplinary Scientific Data as a Trustworthy Digital Repository, by Downs and Chen, which was given during the Fourth International Conference on Digital Repositories, held in Atlanta, Georgia, on May 18 – 21, 2009. The authors very much appreciate advice received from members of the SEDAC Long-Term Archive Board and gratefully acknowledge support received from NASA under contract NNG08HZ11C for work reported here. The opinions expressed are those of the authors and do not necessarily represent those of NASA, the Columbia University Libraries, or Columbia University.

6. References

- Ambacher, B. (2007). "Government Archives and the Digital Repository Audit Checklist". *Journal of Digital Information*, Vol. 8, No. 2. Accessed on September 12, 2009 from <https://journals.tdl.org/jodi/article/view/190>
- Arms, W. Y., Calimlim, M., & Walle, L. (2009). "EScience in Practice: Lessons from the Cornell Web Lab". *D-Lib Magazine*, Vol. 15, No. 5/6. doi:10.1045/may2009-arms
- Caplan, P. (2008). "Repository to Repository Transfer of Enriched Archival Information Packages". *D-Lib Magazine*, Vol. 14, No. 11-12. doi:10.1045/november2008-caplan
- Cocker, M. D. (2005). "Geological Data and Collections in Peril: Case Example in Georgia". *Digital Mapping Techniques '05 — Workshop Proceedings*. U. S. Geological Survey Open-File Report 2005-1428. Accessed on September 14, 2009 from <http://pubs.usgs.gov/of/2005/1428/cocker/index.html>
- Columbia University Libraries. (2006). *Strategic Plan 2006-2009*. Accessed on September 14, 2009 from http://www.columbia.edu/cu/lweb/img/assets/6675/strategicplan_2002-2009.pdf
- Consultative Committee for Space Data Systems. (2004). *Producer-Archive Interface Methodology Abstract Standard*. (CCSDS 651.0-B-1). Adopted as: Space data and information transfer systems -- Producer-archive interface -- Methodology abstract standard (ISO 20652:2006). Accessed on September 12, 2009 from <http://public.ccsds.org/publications/archive/651x0b1.pdf>
- Consultative Committee for Space Data Systems (CCSDS). (2002). *Reference Model for an Open Archival Information System (OAIS)*. Adopted as: Space data and information transfer systems - Open archival information system - Reference model (ISO 14721:2003). Accessed on September 12, 2009 from <http://public.ccsds.org/publications/archive/650x0b1.pdf>
- Downs, R.R., and R.S. Chen. (2003). "Cooperative design, development, and management of interdisciplinary data to support the global environmental change research community." *Science & Technology Libraries*, Vol. 23, No. 4, 5-20.

- Downs, R. R., R. S. Chen, W. C. Lenhardt, W. Bourne, and D. Millman. (2007). "Cooperative management of a long-term archive of heterogeneous scientific data". In *Proceedings of Ensuring the Long-Term Preservation and Value Adding to Scientific and Technical Data (PV 2007)*. Oberpfaffenhofen/Munich, Germany. October 9–11. 2007. Accessed on September 12, 2009 from http://www.pv2007.dlr.de/Papers/Downs_CooperativeManagementOfALongTermArchive.pdf
- Gladney, H. M. (2006). "Principles for digital preservation". *Communications of the ACM*, Vol. 49, No. 2, 111 - 116.
- Groenewegen, D. & Treloar, A. (2008), "A Consortial Institutional Repository Solution, Combining Open Source and Proprietary Software". *OCLC Systems & Services: International Digital Library Perspectives*, Vol. 24, No. 1, 30-39.
- Janée, G., Frew, J., & Moore, T. (2009). "Relay-supporting Archives: Requirements and Progress". *International Journal of Digital Curation*, Vol. 4, No. 1, 57-70. Accessed on September 12, 2009 from <http://www.ijdc.net/index.php/ijdc/article/view/102/77>
- Lagoze, C., Payette, S., Shin, E., & Wilper, C. (2006). "Fedora: An Architecture for Complex Objects and their Relationships". *International Journal on Digital Libraries*, Vol. 6, No. 2, 124-138. doi:10.1007/s00799-005-0130-3 Accessed on September 12, 2009 from <http://arxiv.org/ftp/cs/papers/0501/0501012.pdf>
- Lesk, M. (2008). "Recycling Information: Science Through Data Mining". *International Journal of Digital Curation*, Vol. 3, No. 1, 154-157.
- Littman, J. (2009). "A Set of Transfer-Related Services". *D-Lib Magazine*. Vol. 15, No. 1/2. doi:10.1045/january2009-littman
- Marill, J. L. & Luczak, E. C. (2009). "Evaluation of Digital Repository Software at the National Library of Medicine". *D-Lib Magazine*, Vol. 15, No. 5/6. doi:10.1045/may2009-marill
- Marshall, C. C., Bly, S., and Brun-Cottan, F. (2006). "The Long Term Fate of Our Personal Digital Belongings: Toward a Service Model for Personal Archives." *Proceedings of Archiving 2006*. Springfield, VA: Society for Imaging Science and Technology, pp. 25-30. Accessed on September 14, 2009 from <http://arxiv.org/abs/0704.3653>

Moore, R. (2008). "Towards a Theory of Digital Preservation". *International Journal of Digital Curation*, Vol. 3, No. 1, 63-75.

National Science Board. (2005). *Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century*. National Science Foundation. Accessed on September 9, 2009 from <http://www.nsf.gov/pubs/2005/nsb0540/>

National Science Foundation Cyberinfrastructure Council. (2007). *Cyberinfrastructure Vision for 21st Century Discovery*. National Science Foundation. Accessed on September 9, 2009 from <http://www.nsf.gov/pubs/2007/nsf0728/>

OCLC and CRL. (2007). *Trustworthy Repositories Audit & Certification: Criteria and Checklist*. Accessed on September 12, 2009 from <http://bibpurl.oclc.org/web/16712>