The International Journal of Digital Curation
Issue 2, Volume 2 | 2007

Digital Data Preservation for Scholarly Publications in Astronomy

Sayeed Choudhury, Tim DiLauro, Alex Szalay, Ethan Vishniac,
The Johns Hopkins University

Robert Hanisch,
Space Telescope Science Institute

Julie Steffen,
The University of Chicago Press

Robert Milkey,
American Astronomical Society

Teresa Ehling,
Cornell University

Ray Plante,
University of Illinois/National Center for Supercomputing Applications

November 2007

Abstract

Astronomy is similar to other scientific disciplines in that scholarly publication relies on the presentation and interpretation of data. But although astronomy now has archives for its primary research telescopes and associated surveys, the highly processed data that is presented in the peer-reviewed journals and is the basis for final analysis and interpretation is generally not archived and has no permanent repository. We have initiated a project whose goal is to implement an end-to-end prototype system which, through a partnership of a professional society, that society's scholarly publications/publishers, research libraries, and an information technology substrate provided by the Virtual Observatory, will capture high-level digital data as part of the publication process and establish a distributed network of curated, permanent data repositories. The data in this network will be accessible through the research journals, astronomy data centers, and Virtual Observatory data discovery portals.

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

A fundamental aspect of scientific scholarly communication is the presentation of data. In astronomy and many other fields, data is presented in graphical form (scatter plots, contour maps, gray-scale and color images), but the digital data underlying these graphical representations is not systematically captured and archived. As a result, researchers cannot cite or examine the data from which measurements have been made and conclusions have been drawn. Without this ability, the very essence of the scientific method, with its requirement of validating results, becomes compromised. Astronomers are producing and analyzing data at ever more prodigious rates, with projects such as the Sloan Digital Sky Survey[1] having gathered data at unprecedented rates, raising new challenges and opportunities. This explosion in data-driven science has led to fundamental changes in practice and modes of inquiry, prompting the US-based National Science Foundation (NSF), the UK-based Joint Information Systems Committee (JISC), and other agencies and organizations to advance the evaluation and development of Cyberinfrastructure to support large-scale digital science projects. The Library of Congress National Digital Information Infrastructure and Preservation Program (NDIIPP) and the NSF Blue-Ribbon Panel on Cyberinfrastructure report (National Science Foundation, 2003) stress the essential aspect of digital archiving of datasets to ensure long-term access. Most recently, the US-based Institute of Museum and Library Services (IMLS) awarded this project team a grant through its National Leadership Grant Program, which included a call to "develop pilot projects or programs in data curation." Without immediate action, we may find ourselves in a "digital dark age" losing important, scholarly resources from the scientific domain.

In the United States, NASA's Great Observatories, ground-based national observatories, and major survey projects have archive and data distribution systems in place to manage their standard data products, and these are now interlinked through the protocols and metadata standards agreed upon in the Virtual Observatory. The National Virtual Observatory (NVO) project is playing a leadership role in building services for the astronomy community to access and analyze astronomical data. Through the Virtual Observatory's standard interfaces for data discovery and data access, astronomers, students, educators, and the public will have access to digital astronomy data from around the world. For good reason, the NVO is often cited as one of the quintessential cyberinfrastructure projects. With projects such as NVO, the astronomy community has moved into the forefront of data-intensive digital science, providing a path for other disciplines to consider. However, thus far the scope of the NVO has deliberately not included long-term data curation, focusing instead on data location and data access standards and protocols.

Based on extensive, ongoing dialogue and communication, the NVO project team, led by researchers at Johns Hopkins University (JHU), the California Institute of Technology, and the Space Telescope Science Institute, has concluded that academic research libraries, given their expertise together with long-term, sustainable support from universities, represent the ideal home for long-term preservation and curation of large-scale datasets to support persistent access and scholarly communication. Libraries play an important role in curation and preservation, a role that becomes more complicated with the increasing use of electronic means of content delivery. Digital

[1] http://www.sdss.org

data products require standardized metadata descriptions in order to be scientifically useful and locatable through multiple access portals.

Our vision is to create a cooperative system amongst astronomical researchers, professional societies, publishers, editors, libraries, and the Virtual Observatory in which the literature, the associated digital data, and the underlying data archives interoperate seamlessly and where high-level data products are assured of long-term preservation and curation. This cooperative system involving several large research institutions could serve as a model for other data-driven research disciplines.

Wider dissemination of research results in the age of electronic publishing has increased both the speed and effectiveness of subsequent research in many scholarly disciplines. The data underlying much of this research are more likely than ever before to be born digital. Many research funding bodies and much of the academic community require that both research findings and original data be accessible; this implies a responsibility for integrated data preservation and curation that has only very recently been explored. Technological solutions to large-scale data storage, preservation, and curation are now available through systems such as Fedora[2] and DSpace[3]. National research facilities now provide reasonably comprehensive data archives. What is missing is an end-to-end process for capturing, curating, preserving, and providing long-term access to the highly processed data described in peer-reviewed publications. We are prototyping such a process. Our goals include assessing the scientific impact of this new approach to astronomical data, as well as working out sustainable business models for increasing the value of data in this way.

Our prototype phase focuses on astronomy because of the technological maturity of electronic publications and data management in this discipline, and because of the wide access to digital data archives. Astronomers acquire, calibrate, and analyze hundreds of terabytes of data from both ground- and space-based observatories every year. Raw and processed data mainly reside in observatory archives and data centers, and are available (after nominal proprietary periods) for re-use by any researchers. However, the most refined digital data products - the recalibrated and specially processed digital images, spectra, and time series that are graphically shown in the peer-reviewed literature - are not systematically preserved. As a result, assertions and analyses based on these data cannot be independently verified, and creative re-use of the data (e.g., to examine other objects in the same field of view, or to combine multiple datasets in correlative studies) is possible only through private collaborations.

The Objective of the Virtual Observatory

The objective of the Virtual Observatory[4] is to support new science by greatly enhancing access to data and computing resources. The VO is intended to make it easy to locate, retrieve, and analyze data from archives and catalogs worldwide, and it assumes that astronomical data is distributed rather than centralized. Thus, the VO is concerned with data discovery, data access, and data integration, capabilities that are enabled through the use of standard protocols for registering the existence and location of data and for requesting data that satisfies the user's interests. These standards are

[2] http://www.fedora.info

[3] http://www.dspace.org

[4] http://us-vo.org and http://ivoa.net

developed on a national (the US National Virtual Observatory) and international (the International Virtual Observatory Alliance) basis, with the goal of having one set of standards that is accepted worldwide.

As is illustrated in Figure 1 below, the VO provides common discovery and access methods so that users of astronomical resources need not interact with them one at a time, each with a different user interface and peculiar data formatting, but rather through a common interface. The VO interoperability standards allow for diverse resources to be queried in a standard way, and for responses from data providers to be encapsulated with standard descriptive information - metadata - sufficient for a web-based portal or other application to be able to understand the underlying data.

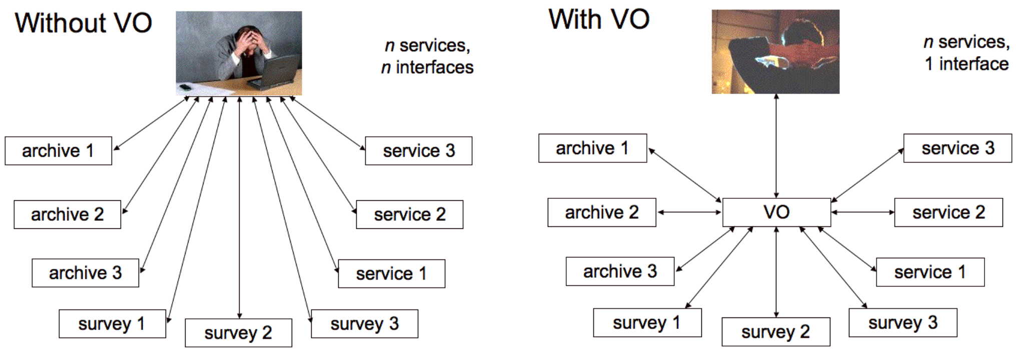

Figure 1 *Left*: Astronomy data collections and services as seen by the astronomer, without the Virtual Observatory. Each archive, survey, or service has its own interface and must be discovered by the end user. *Right*: Astronomy data collections and services as seen by the astronomer, using the Virtual Observatory. Resources with VO-compliant interfaces are accessible through a single portal or application.

The essence of the Virtual Observatory is *interoperability*. Data discovery, data access, and database queries are enabled by metadata standards. An open, service-oriented architecture allows for distributed applications with web services wrappers to interact and exchange data. Data storage is inherently distributed, both for institutionally supported data collections and for user-generated data. Heavily used and/or large-scale datasets are replicated to assure availability and proximity to large-scale computational resources. Access to proprietary data or controlled-access computational facilities is controlled through authentication and authorization services based on standard Grid protocols.

A primary goal of the Virtual Observatory is to provide integrated access to archival data and derived data products: catalogs, tables, and highly processed images, spectra, and time series. The initial focus of NVO development has been on providing access to the former - the archival datasets that are already available via public interfaces on the web, but often with unique and incompatible interfaces. Derived data products are the purview of either dedicated large projects (such as the 2MASS or SDSS sky surveys) or of individual researchers. Large projects have thus far worked to provide access to these high-level products. The valuable data from individuals or small collaborations sometimes appear in the electronic journals and sometimes on personal web sites, but most often these data are not available at all in any standard form or via any standard interface.

The Data Preservation Problem

What does data curation mean for astronomers? Astronomers' data include images, spectra, catalogs/tables, and documents or free form data. Individual or teams of astronomers using ground-based and space-based instruments capture or create a portion of these data that provide the foundation for research and publication. Most of these data reside in systems optimized for the storage and retrieval of data from particular telescopes or facilities, but lacking generic data access or query mechanisms (a situation the NVO is beginning to rectify). In any case, major astronomy archives focus on standard data products, and less so on highly processed images or spectra that are associated with peer-reviewed publications.

Astronomers publish their research results in conference proceedings, electronic preprints, PhD theses, and peer-reviewed journals. The journals are the primary long-term repository for scholarly research, and in astronomy in particular, journal articles often contain and describe highly processed data. However, the journals typically only contain graphical representations of data: pictures rather than digital images, and graphs rather than the numeric data represented in the graph. In recent years astronomy journals have worked in partnership with data centers, most notably the CDS in Strasbourg[5], to obtain from authors, review, and retain tabular data. Machine-readable tables are now frequent adjuncts to published papers, especially for large catalogs that used to be relegated to the journal supplement series.

However, journals typically balk at author requests to accept digital imaging data, such as the FITS[6] files that underlie the graphical representations (gray scale and color images, contour plots, polarization maps) that appear in the journals. And indeed, some authors balk at the idea of making their digital data available to others to examine because they are not finished with their analysis. Yet readers of the journals increasingly want to be able to compare the results of others with their own work, to bring together data from diverse sources, and to undertake investigations that synthesize information. Scientists willing to share their data often find that, some years later, they neither know where their data are nor whether the media hosting them remains generally readable. If the digital data underlying a research paper are not available for inspection, then the results and interpretation cannot be independently verified and we do a disservice to the scientific process.

Without an integrated system for individual scientists to deposit these data into a persistent library-based archive or NVO-standard interfaces for access, it is impossible to query these data, identify gaps in knowledge, cite the data within publications, or preserve them for long-term access. The (primarily) public funds invested in data collection do not provide maximum scientific return. The barrier for participation in such a system should be low. Ideally, individual scientists should be able to register their processed, high-level data with library archives as easily as they can create web pages.

[5] http://cdsweb.u-strasbg.fr/

[6] http://fits.gsfc.nasa.gov/

Architectural Components

A high-level picture of the digital data preservation architecture is shown in Figure 2. Starting at the top, we have the publication and editorial process, which includes data capture, metadata tagging, and linking between the digital data and journal article using persistent identifiers. Data ingest is handled by a data storage appliance, a black-box storage system whose functions are limited to storage of digital data objects and associated metadata databases. Data storage appliances can be hosted by the journal publisher, scientific society, research libraries, and at VO sites, and various sites can host both astronomy data and other content as they desire. The astronomy data will be replicated among the storage appliances for redundancy, and the "VOSpace" environment for distributed storage provides, as necessary, authentication and authorization services for datasets that may have restricted access such as proprietary periods.[7] Data access is provided directly from the journal portals, or through the VO once the digital data collections are registered with the VO resource registry. Research libraries and journal editorial staff are involved with metadata curation through direct interaction with the metadata database.

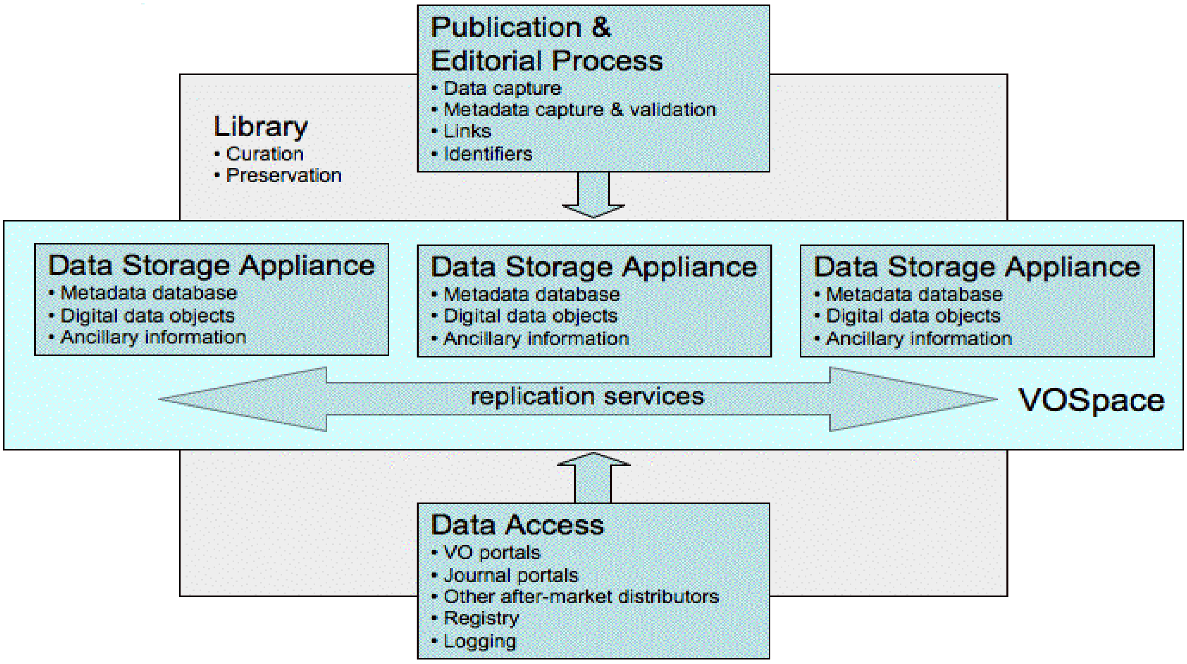

Figure 2. Architectural components of the digital preservation facility.

In the astronomy community there is a long-established partnership between the dominant, non-profit publishers such as the American Astronomical Society and its production partner the University of Chicago Press (*Astrophysical Journal*, *Astronomical Journal*) and astronomy data centers and bibliographic services (Astrophysics Data System, ADS, in the US; Centre de Données astronomique de Strasbourg, CDS, in Europe). Our initiative offers an opportunity to move libraries from the periphery of projects such as NVO to the center of digital archiving and data curation efforts, and to establish a three-fold collaboration—publishers, an association of libraries, and NVO - that assures universal and long-term access and preservation.

[7] http://www.ivoa.net/twiki/bin/view/IVOA/IvoaGridAndWebServices

By incorporating NVO web services into a Fedora digital library framework we will provide mechanisms for long-term digital archiving of astronomical tables, catalogs, spectra, images and documents that facilitate data publishing for astronomers and scholarly journals.

A Prototype Project

Our near-term goal is to implement an end-to-end prototype digital data preservation facility using astronomy scholarly publications as a test-bed. Astronomy is an ideal discipline to start with, as essentially all digital imaging data is available in a single standard format (FITS), the VO projects are defining further standards for interoperability for catalogs and spectral data, the community is small and highly e-publishing-aware, and there are few restrictions on data access and exchange.[8]

In the astronomy community there is already some experience with digital data preservation. One of our team members, Ray Plante (NCSA), developed the Astronomy Digital Image Library (ADIL) ten years ago (Plante, Crutcher, & Sharpe, 1996). ADIL allows researchers to upload digital data collections and to link them with the peer-reviewed publication that describes them. ADIL, however, has not been used very extensively, and we believe that one of the main reasons for this is because the data upload and curation process is totally separate from the publication process. Thus, our idea is to draw on the strengths of ADIL, but to integrate digital data management into the publication process. This includes data capture, review, metadata tagging and validation, and storage.

We will develop a set of web services that link literature and reference materials to astronomical datasets. These services will reflect actual use cases as defined by the NVO. For example, “identify all literature and images within this portion of the sky.” We will map these web services into a Fedora-based digital library framework. Deposited data will ultimately be archived within the Library, which will serve as a digital annex for publications. Fedora's ability to integrate web services and object discovery facilities are particularly relevant in this regard. Through this effort, we will develop an appliance that will comprise both the hardware and software to manage this service. This appliance will be installed at our partner libraries at the University of Edinburgh and the University of Washington.

Fedora (Payette, Staples, & Wayland, 2003) is an open source repository system being actively developed by the University of Virginia and Cornell University. The fact that Fedora is open source will give us the ability to modify the system, as necessary, and to distribute the appliance without concerns about software copyright. Unlike some other repository applications, Fedora was designed to support the association of behaviors with the digital objects it contains. These associations are called “disseminations” in Fedora. So, instead of simply returning the content, as it was stored into the repository, Fedora can render the content or act on it in different ways. This coupling of digital objects with behaviors will allow richer interaction with the deposited content.

[8] As our collaborator Jim Gray of Microsoft liked to say, ”Astronomy data is worthless!” By which he meant it has no commercial value and thus experimenting with it comes without concerns for licenses, fees, etc.

Ingesting content into Fedora can support digital preservation in a number of ways:

- Fedora supports the typing of datastreams that can support format migration. These datastreams can be validated using checksums and applications like JHOVE and the Unix “file” command.
- Fedora’s relationship database (the RDF triple store) allows for the tracking of relationships between objects. This feature could be useful for building preservation packages for standard Fedora digital objects and for the “network overlay” digital objects that are built by linking together parts of other objects.
- In upcoming versions, Fedora will support the concept of content models, which brings semantics to the management of digital objects. Content models will allow higher levels of object verification and validation (above and beyond validation of individual datastreams).
- Fedora supports replication, which can be used to duplicate information from one repository to another, allowing digital objects to survive disasters or failures in a single location.
- While Fedora uses a database to support its operation, a Fedora repository can be completely rebuilt from just the contents of its filesystem.

In addition to providing a default dissemination based on content type (e.g., image, spectra, catalog), we will use these facilities to link deposited content with appropriate NVO services. These interfaces will allow users of a digital object to be drawn into a rich interactive experience that uses data from the selected object as the starting point for further discovery.

For example, the image content (TIFF, JPG, etc.) of an image digital object might be displayed to the end user with links to a web service that allows interaction with a larger portion of the sky. Data from the object would be passed as parameters to the service, allowing the service to focus on the correct portion of the sky and, perhaps, to highlight appropriate objects.

Data discovery and access will be possible through the journal portals (from the research papers themselves) and VO portals (which provide more open-ended queries, such as “find me all images of NGC 4151”, or “where are there spectra of AGNs?”). Data management and curation will incorporate the expertise and guidance of the research libraries, whose organizations are committed to long-term preservation and whose experience in cataloging and metadata quality checking will be invaluable. As a result of providing a simple mechanism for researchers to upload, register, and annotate their processed data in a permanent digital archive, we enable the VO to integrate the collection of processed data products into the general framework for data discovery and access. The standard VO methods for data access (catalogs, spectra, images) can be implemented as front-end services to the Fedora-based, distributed collection of high-level data products.

A somewhat larger challenge comes in the area of data discovery, which depends on the availability of reliable metadata describing the datasets in a collection. To support search and discovery, technical and descriptive metadata will be extracted from deposited objects and mapped to one or more common metadata formats. While the details will be developed over the course of the project, it is anticipated that these formats will include simple Dublin Core[9] and a format designed to support information specific to the astronomy domain. We will draw upon the VO experience in defining metadata standards for resource discovery and associated data models. Metadata will be made available using the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH). [10] Content will be harvested frequently so that new data quickly become available to the astronomy community.

With the cooperation of the American Astronomical Society, the editorial staff of the *Astrophysical Journal* and *Astronomical Journal*, and the Institute of Physics, we will work with publishers to define suitable data submission formats and procedures. Since the association of libraries will manage and preserve the deposited datasets, we will reduce the participation and entry barrier for publishers. Through these combined efforts, we will create a fully integrated network of processes, tools and systems to support ingestion and preservation of processed datasets to support publication.

Background Research

The Digital Knowledge Center (DKC), now relaunched as the Digital Research and Curation Center (DRCC), at the Sheridan Libraries of JHU represents a unique organization focused on digital library research and development. While the DRCC is housed physically and administratively within the Libraries, its staff includes individuals with backgrounds in computer science, engineering, mathematics and cognitive science. This combination of perspectives has resulted in a comprehensive, diverse approach to digital library development. DKC projects have focused on digital workflow management, especially the ingestion of and access to large digital collections. Through existing grants, the DKC has built both hardware (Suthakorn, Sangyoon, Zhou, Choudhury, & Chirikjian, 2003) and software (Droettboom et al., 2002), including tools for automated metadata generation (DiLauro, Choudhury, Patton, Warner, & Brown, 2001).

More recently, the DKC has focused on repository research, especially as it relates to repositories' ability to support a range of services, especially digital preservation. Led by the DKC, JHU participated in the Archive Ingest and Handling Test (AIHT), a program within the Library of Congress' NDIIPP framework (DiLauro, Patton, Reynolds, & Choudhury, 2005). The AIHT provided practical experience with ingestion of an archive that features multiple file formats, and varying levels of metadata. The AIHT test provided an excellent opportunity to test the capabilities of existing repository systems such as DSpace (Smith et al., 2003) and Fedora. JHU was the only institution that evaluated multiple systems. Through a grant from the Mellon Foundation, JHU has conducted a comprehensive and diverse technology analysis of repositories and services. [11]

In addition to these detailed explorations of repositories, the DKC is also involved in two UK-based repository projects funded by JISC. Through one of these JISC projects, STORE[12], the project team has further defined and articulated the specific needs of astronomers and other scientists for data curation, especially as it relates to

[9] http://dublincore.org/

[10] http://www.openarchives.org/OAI/openarchivesprotocol.html

[11] https://wiki.library.jhu.edu/display/RepoAnalysis/ProjectRepository

[12] http://jiscstore.jot.com/WikiHome (See also Graham Pryor's article in *IJDC* Vol.2,*(1)*)

electronic publications. JISC has coordinated its Digital Repository Programme training and development with JHU's Mellon-funded repository analysis, and invited PI Choudhury to two meetings in the UK and the Netherlands. Finally, the DKC evaluated LOCKSS (Reich & Rosenthal, 2001), and continues to manage a LOCKSS appliance. Through our extensive, rigorous and objective evaluation, we have concluded that Fedora represents the best system for this particular effort.

Conclusions

By implementing a prototype we hope to understand operational costs and thus be able to develop a long-term business plan for the preservation of peer-reviewed journal content and the associated supporting data. The availability of such facilities for digital data preservation will undoubtedly lead to changes in policies affecting data access. Peer pressure may initially encourage researchers to contribute their data to the repository, but in time such contributions might become mandatory. It seems likely that published papers having digital data available will be more heavily used, and thus more heavily cited, than papers lacking such data. In the longer term it will be important to evaluate the impact on scientific productivity through citation analyses and community feedback.

Acknowledgements

JISC, through Project STORE, has provided funds to support this work. IMLS has recently announced an award to support this work through its National Leadership Grant Program. Therefore this work is supported through The Institute of Museum and Library Services National Leadership Grant LG0606018206. The Scholarly Publishing and Academic Resources Coalition (SPARC), Microsoft, TeraGrid, and NVO have also committed contributions to support this work. Andrew Lawrence, head of the Astrogrid Project, provided valuable support in developing the IMLS proposal.

References

DiLauro, T., Choudhury, G. S., Patton, M., Warner, J. W. & Brown, E. W. (2001, April). Automated name authority control and enhanced searching in the Levy Collection. *D-Lib Magazine 7,*(4). Retrieved September 30, 2006, from http://www.dlib.org/dlib/april01/dilauro/04dilauro.html

DiLauro, T., Patton, M., Reynolds D., & Choudhury, G.S. (2005, December). The archive ingest and handling test: The Johns Hopkins University report. *D-Lib Magazine 11,*(12). Retrieved September 30, 2006, from http://www.dlib.org/dlib/december05/choudhury/12choudhury.html

Droettboom, M., Fujinaga I., MacMillan K., Choudhury G. S., DiLauro T., Patton M., & Anderson T. (2002). Using the GAMERA framework for the recognition of cultural heritage materials. *JCDL 2002: The Second ACM+IEEE Joint Conference on Digital Libraries: 11-17*. Portland: ACM Press. Retrieved September 30, 2006, from http://portal.acm.org/citation.cfm?id=544220

National Science Foundation. (2003). *Report of the National Science Foundation Blue-Ribbon Advisory Panel on cyberinfrastructure*. Retrieved September 30, 2006, from http://www.nsf.gov/publications/pub_summ.jsp?ods_key=cise051203

Payette, S., Staples, T., & Wayland, R. (2003, April). The Fedora Project: An Open-source digital object repository management system. *D-Lib Magazine 9,*(4). Retrieved September 30, 2006, from http://www.dlib.org/dlib/april03/staples/04staples.html

Plante, R. L., Crutcher, R. M., & Sharpe, R. K. (1996). ADASS V. In G. H. Jacoby & J. Barnes (Eds.), *ASP Conf. Ser. 101,* 581.

Reich, V., & Rosenthal, D. S. H. (2001, June). LOCKSS: A permanent web publishing and access system. *D-Lib Magazine 7,*(6). Retrieved September 30, 2006, from http://www.dlib.org/dlib/june01/reich/06reich.html

Smith, M., Barton, M., Bass, M., Branschofsky, M., McClellan, G., Stuve, D., et al. (2003, January). DSpace: An open source dynamic digital repository. *D-Lib Magazine 9,*(1). Retrieved September 30, 2006, from http://www.dlib.org/dlib/january03/smith/01smith.html

Suthakorn, J., Sangyoon, L., Zhou, Y., Choudhury, G. S. & Chirikjian, G. S. (2003, July). An enhanced robotic library system for an off-site shelving facility. In *International Conference on Field and Service Robotics*.