Contexts and Contributions:
Building the Distributed Library

Martha L. Brogan

Digital Library Federation
Washington, D.C.
2006

Additional print copies are available for 35.00 USD per copy. Orders must be placed through DLF's Web site.

∞ The paper in this publication meets the minimum requirements of the American National Standards for Information Sciences—Permanence of Paper for Printed Library Materials, ANSI Z39.48-1984.

Note the manuscript was submitted for publication on June 5, 2006; final edited version completed October 2, 2006.

TABLE OF CONTENTS

About the Author

Martha L. Brogan is the author of two previous studies commissioned by the Digital Library Federation and the Council on Library and Information Resources:

> *A Survey of Digital Library Aggregation Services* (DLF, 2003)
> http://www.diglib.org/pubs/brogan/
>
> *A Kaleidoscope of Digital American Literature* (DLF and CLIR, 2005).
> http://www.diglib.org/pubs/brogan0505/

Ms. Brogan is an independent library consultant with two decades of experience in research libraries at the University of Minnesota, Yale University, and Indiana University, where she served as associate dean and director of collection development from 1998 to 2003. She currently holds an appointment on CLIR's Scholarly Communications Advisory Committee. In 2001 Ms. Brogan participated as a fellow in the Frye Leadership Institute sponsored by CLIR, Educause, and Emory University.

Acknowledgments

Preface

Martha L. Brogan's *Contexts and Contributions: Building the Distributed Library* is a major contribution to the Digital Library Federation's (DLF) suite of work that focuses on the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH). With generous funding from the Institute of Museum and Library Services, DLF has harnessed deep OAI expertise from the University of Michigan, the University of Illinois at Urbana-Champaign, and Emory University to prototype "next-generation" OAI services informed by advisory panels of scholars and technical experts; to build registries of providers to aid in the creation of new OAI-based services; and to formulate best practices for sharable metadata that focus what we have learned collectively for innovative library services. The best practices work has received intellectual and practical support from our colleagues at the National Science Digital Library (NSDL), a service of the National Science Foundation (NSF).

Contexts and Contributions had its starting point in a 2003 survey of digital library aggregation services compiled by Martha Brogan for DLF: *A Survey of Digital Library Aggregation Services* < http://www.diglib.org/pubs/brogan/>. This environmental scan was influential in the understanding of our early attempts to craft aggregated digital library services that served students and scholars well, and it had a very positive impact on the development of the services that followed.

The current work is more difficult because the environment is maturing, and changing rapidly. Its value and timeliness is increased because of that, and I am proud that DLF can sponsor such a detailed evaluation of a shifting, but critically important landscape. Martha Brogan's current study draws our attention to "major developments affecting the ecosystem of scholarly communications and digital libraries" and gives us all a rich comparative analysis of digital library aggregation services, including a clear-sighted view of—in Martha's words—"the obstacles requiring further attention to realize … an open, distributed digital library."

The Digital Library Federation is delighted to acknowledge our funders and expert partners in this important work. We are pleased to have another opportunity to underscore our commitment to those standards, tools, and technologies that allow us to build innovative services that scholars and students need to produce richer teaching, learning, and scholarship.

David Seaman
Executive Director
Digital Library Federation
October 2006

PART I: INTRODUCTION

1.0   Laying the Foundation

Since its founding ten years ago, the Digital Library Federation (DLF) aims to advance the goal of deep sharing of academic digital resources and services. It creates and promotes standards and strategies that will lead to an extensive, open, distributed digital library with coherent pathways for scholars to discover, access, and use meaningful content. Executive Director, David Seaman, refers to DLF's twin goals of achieving "mass and malleability" through federated content and interdependent services.

The DLF, along with the Coalition for Networked Information (CNI) and the National Science Foundation (NSF), co-funded the early development of the OAI protocol (Open Archives Initiative Protocol for Metadata Harvesting or OAI-PMH) as a low-barrier means to share metadata and a technical framework to achieve cross-repository interoperability. According to Seaman, the OAI protocol is a key component of DLF's commitment to build finding systems and services that are flexible and useful in various settings for different constituents. He asserts that it provides a practical means to create a collaborative test bed that is larger and more complex than what DLF-member institutions could produce separately. Moreover, it forces developers to focus on building shareable metadata that is useful to others, who may discover a digital object outside its original context, may not share the same assumptions about its most salient characteristics, and may want to use it in new ways (Seaman 2005b).

Since its initial release in 2001, OAI-PMH has become a widely accepted, international harvesting protocol for sharing metadata between services. More than 1,000 OAI-compliant archives, representing a wide variety of domains and institutions, are operational in over 40 countries. OAI functionality is now a standard component of many vendor's integrated library systems and repository services. More recently, OAI's principal developers are exploring enhanced digital library systems and Web services that, among other features, could manage the transfer of diverse content as well as metadata (Lagoze et al. 2006a, Van de Sompel et al. 2004 and 2005). [1] As Seaman and others point out, scholars need to be able to do more than view digital content in the

---

[1] An invitational meeting convened by The Andrew W. Mellon Foundation in April 2006 aimed to reach agreement "on the nature and characteristics of a limited set of core, protocol-based repository interfaces (REST-full and/or SOAP-based Web services) that allow downstream applications to interact with heterogeneous repositories in an efficient and consistent manner" (Hey et al. 2006). Documentation about the meeting and follow-up activities are available from http://msc.mellon.org/Meetings/Interop/follow-up/jcdlpanel_abstract.pdf .

context of its original producer. Digital infrastructures must allow scholars to bring content into their own environment where they can apply discipline-specific tools of inquiry. As the digital content in repositories proliferates, efficient and consistent interoperability specifications are essential for effective downstream applications across a full spectrum of scholarly information arenas extending from e-research and e-learning to Web publishing and administrative computing (Hey et al. 2006; McLean and Lynch, 2004).

1.1   Aim, Scope, and Methodology

This report updates and expands on "A Survey of Digital Library Aggregation Services," originally commissioned by the DLF as an internal report in summer 2003, and released to the public later that year. It highlights major developments affecting the ecosystem of scholarly communications and digital libraries since the last survey and provides an analysis of "OAI implementation demographics," based on a comparative review of repository registries and cross-archive search services. Secondly, it reviews the state-of-practice for a cohort of digital library aggregation services, grouping them in the context of the "problem space" to which they most closely adhere. Based in part on responses collected in fall 2005 from an online survey distributed to the original core services, the report investigates the purpose, function and challenges of next-generation aggregation services. On a case-by-case basis, the advances in each service are of interest in isolation from each other, but the report also attempts to situate these services in a larger context and to understand how they fit into a multi-dimensional and interdependent ecosystem supporting the worldwide community of scholars. Finally, the report summarizes the contributions of these services thus far and identifies obstacles requiring further attention to realize the goal of an open, distributed digital library system.

The new report aims to inform DLF's continuing efforts "to foster better teaching and scholarship through easier, more relevant discovery of digital resources, and a much greater ability for libraries to build more responsive local services on top of a distributed metadata platform," as articulated in its successful IMLS National Leadership Grant, "The Distributed Library: OAI for Digital Library Aggregation."[2] Extending over a two-year period from October 2004 through September 2006, the grant enables DLF to prototype a "second generation" OAI finding system. Concurrently, it affords DLF the opportunity to address challenges identified in the 2003 survey and voiced by early OAI adopters. In particular, DLF is building a comprehensive OAI registry, establishing best practices for shareable metadata, improving communication between data and service providers, and developing curricular materials and training sessions to introduce OAI best practices to a widening circle of institutions (Shreeves et al. 2005).

---

[2] Documents, presentations, and a timeline of milestones are available from the DLF's Web site, http://www.diglib.org/architectures/oai/imls2004/.

Using the 2003 survey as a point of departure, this companion report takes a fresh look at the evolution of interoperability and federating heterogeneous content, especially as realized through implementation of the OAI protocol. It re-examines the original set of digital library aggregation services as well as representative new initiatives in an effort to identify trends—progress, needs, and challenges. How are they evolving over time? What have they achieved? What is impeding their progress? How do they envision their future? An online survey conducted in fall 2005 gathered baseline information from more than forty aggregators. As recorded in the series of Update Tables which appear throughout this report, the questionnaire inquired about the services':

- Organizational model
- Subject
- Function
- Primary Audience
- Status
- Size
- Use
- Accomplishments
- Challenges
- Tools or Resources Needed
- Goals of the "Next Generation" Resource

The resource descriptions in section 4.0 also reflect the author's experience in testing out the services, and a selective review of the literature about them. Responses to questions about plans to modify metadata practices to conform to new best practice guidelines promoted by the DLF in collaboration with NSDL, and about whether or not the service is registered with various OAI registries, are also presented in general terms in the report.

Several services responding to the survey proved beyond the scope of the report. Appendix 3 lists the original services discussed in the 2003 report that are no longer included because they fulfilled their mission as experimental or pilot projects, or otherwise ceased operation. In the end, a cohort of 40 services forms the basis of the current study, about one-third of which are new. In addition to these core services, the report points to many other corollary services. Overall, the selection serves as a representative sample of different types of aggregations, focusing mainly on domain or subject-based initiatives in the sciences and humanities. Reflecting OAI-PMH's early application to e-prints, the sample is heavily oriented to text-based aggregations.

The report was prepared over a nine-month period, beginning with survey data collection in September through November 2005 and continuing with a review of services until May 2006. Of course, throughout this period the services continued to evolve and change. The Update Tables are a snapshot from fall 2005, whereas the

individual resource descriptions range in date from late January to mid-May 2006. The majority of project representatives reviewed drafts about their services. Their comments strengthened the report; however, the author bears responsibility for errors or misinterpretations in the final copy.

1.2 Overview of 2003 Findings

The 2003 report examined the baseline features of the aggregations under review, including their organizational model, subject coverage, function, audience, status, and size. The conclusion of this report (section 5.0) includes a discussion of how these features have evolved from 2003 to 2006.

Lacking consensus about how to classify services, the 2003 report grouped them into five categories by function in order to facilitate a review of trends and challenges. Each category evinced a particular set of critical issues, as documented in Table 01.

Table 01: 2003 Critical Issues by Functional Category

FUNCTIONAL CATEGORY	CRITICAL ISSUES
Open Access E-print Archives and Servers • arXiv • NASA Technical Reports Server • PubMed Central	• Gaining momentum through the open access and self-archiving movement but need to attract authors in sufficient numbers to develop repositories of sufficient size to be of interest. • Finding efficient ways to manage copyright issues.
Cross-Archive Search Services and Aggregators • General: Arc, OAIster, Cyclades • Community-Based: NDLTD Union Catalogs, OLAC, Sheet Music Consortium • Subject-Based: UIUC Digital Gateway to Cultural Heritage Materials, Grainger Engineering Library at UIUC, Citebase, Archon	• Having sufficient data to make the service worthwhile to use. • Providing the user with sufficient information so they understand the scope and currency of coverage. o What results are retrieved: links to the source collection-level only, direct links to digital objects, links to analog objects, links to resources available to restricted users? • Providing the user with a "context" in which to understand the items retrieved, i.e. items are detached from their richer original-source native environment. From what original collection is the item derived and how can it be accessed?
From Digital Collections to Digital Library Environments • Cultural Heritage: American Memory, Heritage Colorado	• Organizational sustainability with increasing attention paid to governance structures and the need for business plans. • Management and preservation of data or

• Humanities: The Perseus Digital Library • Sciences: National Science Digital Library, SMETE Digital Library, ENC Online, BiosciEdNet (BEN), DLESE	data "curation"—assigning long-term responsibility. • Managing comprehensive "collections" or "libraries" while providing subsets of users with organized pathways through the content and services tailored to their needs. • Figuring out how to make digital representations reusable for different purposes by different constituents. • Transitioning from digital libraries to digital learning environments, with more attention on users and uses.
From Peer-Reviewed Referratories to Portal Services • Peer-Reviewed Learning Resources: MERLOT • Expert & Machine-Gathered Internet Resources: INFOMINE, UK's Subject Portals • Scholar-Designed Portal: AmericanSouth • Research Library Portals: ARL Scholars Portal, AARLIN Scholars Portal	• Maintaining sustainable systems of quality control in the face of burgeoning resources. • Managing security, user authentication and access. • Linking to the "appropriate copy." • Balancing the competing needs for single-search interfaces with ability to conduct advanced searches. • Rewarding participation by scholars.
Specialized Search Engines • Flashpoint • CiteSeer • Scirus	• Filtering to access quality information. • Offering citation analysis alongside other sophisticated search and retrieval services.

The report then identified three overarching factors that constrained wider use of the resources.

1. The absence of a user-friendly comprehensive registry of OAI-compliant services geared towards users to improve resource discoverability.
2. The lack of priority given to creating and exposing OAI-compliant metadata to meet minimal let alone enhanced standards, coupled with problematic issues of granularity and the need to amass more object-level data.
3. The aggregations did not provide users with a meaningful "context" or match the level of refinement available from the resource's native environment or of their proprietary counterparts.

It concluded by highlighting five future directions to pursue: (1) giving more attention to users and uses; (2) finding solutions to digital rights management and digital content preservation; (3) building personal libraries and collaborative workspaces; (4) putting digital libraries in the classroom and digital objects in the curriculum; and (5) promoting excellence.

The 2006 reexamination reveals significant progress across multiple fronts, while also highlighting some of the fundamental problems that continue to thwart the ambition of achieving large scale, interoperable digital library environments of undisputed importance to scholars.

1.3   Catching-up and Staying Current: A Review of the Literature

This study builds on the 2003 survey and assumes that readers have a basic familiarity with the services, terms and concepts discussed in the original report, including the Open Archives Initiative and the OAI Protocol for Metadata Harvesting (OAI-PMH). If not, Hunter's tutorial on "OAI and OAI-PMH for absolute beginners: a non-technical introduction" delivered at the CERN workshop on Innovations in Scholarly Communication (OAI4) in Geneva, Switzerland in October 2005 provides an excellent overview.

Figure 01: OAI-PMH Structure Model
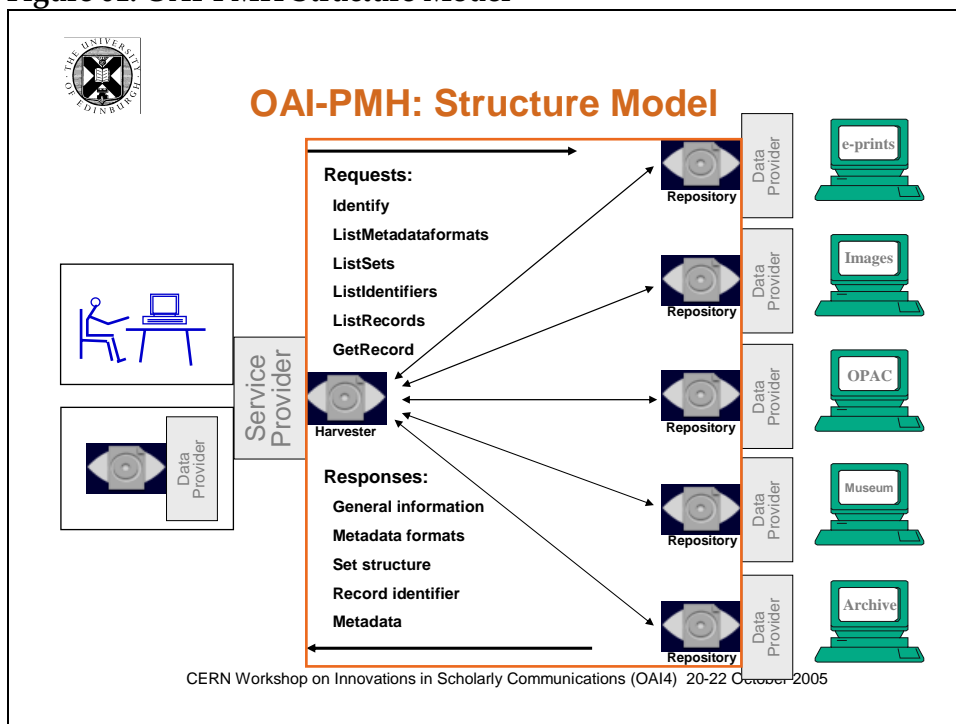
Source: Hunter 2005: http://eprints.rclis.org/archive/00005512/ Used with permission.

Although the report's bibliography is extensive, it is by no means comprehensive. It focuses on articles that have appeared since August 2003. As projects and the field mature, the literature is burgeoning, making it a Sisyphean task to keep up. Fortunately, most of the services under review maintain bibliographies with relevant documents, presentations, and publications at their Web sites. In addition, the sources cited below

are especially helpful for background information and staying current with the wide range of issues covered by this report.

- Peter Suber's Weblog, "Open Access News" (OAN), and its monthly counterpart, "SPARC Open Access Newsletter," are invaluable resources for keeping up with the open access movement. OAN has a search engine, making it easy to retrieve information by event, name or theme. In addition, Suber provides an overview of OA, lists upcoming and past conferences, and maintains a timeline of OA milestones. All are accessible from http://www.earlham.edu/~peters/fos/fosblog.html.
- OCLC's Vice President for Research and Chief Strategist, Lorcan Dempsey's Weblog (also distributed via weekly digests) covers a wide spectrum news and analysis of library issues, services, and networks. Available from http://orweblog.oclc.org/.
- Charles W. Bailey, Jr., *Open Access Bibliography*, published by ARL in 2005 is now freely available online at: http://www.escholarlypub.com/oab/oab.pdf. Bailey's other digital works, including regular updates to his *Scholarly Electronic Publishing Bibliography*, are accessible from http://www.digital-scholarship.com/.
- Three recent books discuss respectively *The Access Principle* (Willinsky 2006), *The Institutional Repository* (Jones et al. 2006) and, *Open Access: Key Strategic, Technical and Economic Aspects*, (Jacobs, ed., 2006).
- The "JISC Disciplinary Differences Report" (Sparks 2005) reviews the literature and reports survey findings related to behaviors, attitudes, and practices within and across disciplines (accessible from http://www.jisc.ac.uk/index.cfm?name=jcie_scg).
- "Use and Users of Digital Resources: A Focus of Undergraduate Education in the Humanities and Social Sciences" (Harley et al. 2006) prepared for the Center for Studies in Higher Education (CSHE), University of California, Berkeley investigates how and if available digital resources are being used in undergraduate teaching (http://cshe.berkeley.edu/research/digitalresourcestudy/report/digitalresourcestudy_final_report_text.pdf). Alan Wolf and Flora McMartin are conducting a similar study about the use of digital resources in science education that will use CSHE's research design to investigate, "Faculty Participation in NSDL— Lowering the Barriers" (NSDL Project ID 435398 awarded January 1, 2005).
- Bettinna Fabos, Issue Editor of "The Commercialized Web: Challenges for Libraries and Democracy," *Library Trends*, Spring 2005. Part II on "Harnessing the Web for Noncommercial Purposes" includes articles about OAI-PMH (Shreeves et al; Liu et al.) as well as related topics on collaboration, portal development, and standardization.
- Distributions lists: about OAI for generalists or implementers from the Open Archives Initiative (http://www.openarchives.org/), about digital repositories from the UK Joint Information Systems Committee (JISC)

(http://www.jiscmail.ac.uk/lists/JISC-REPOSITORIES.html), about digital
libraries from IFLA (http://infoserv.inist.fr/wwsympa.fcgi/info/diglib/), and
about new resources (primarily electronic) and developing trends, with
especially strong coverage about search engines, ResourceShelf
(http://www.resourceshelf.com/).

- *D-Lib Magazine* (http://www.dlib.org/), *Ariadne* (http://www.ariadne.ac.uk/), and
  Cyberinfrastructure Technology Watch Quarterly
  (http://www.ctwatch.org/quarterly/) regularly publish articles and news
  germane to this report.

1.4 Problem Spaces

The diagram below illustrates the overall scholarly information environment described
in the report.

Figure 02: Scholarly Information Environment
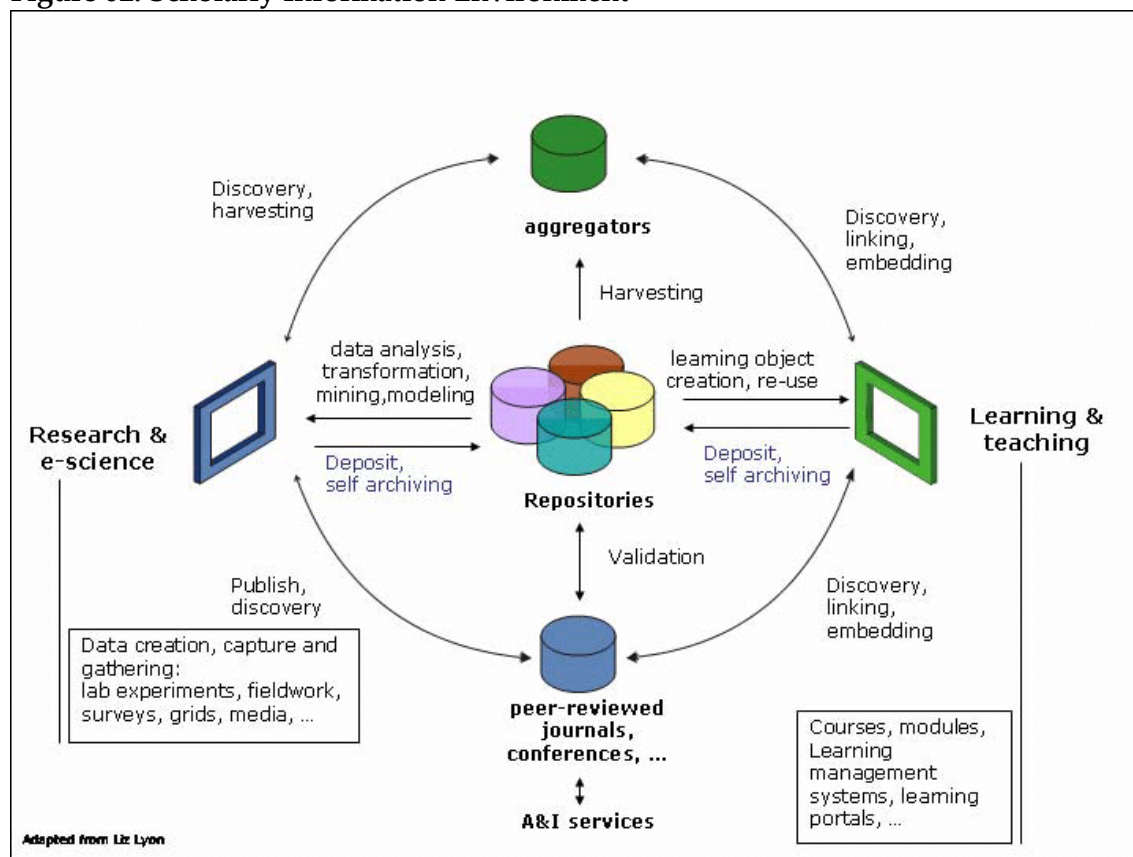
©2006 OCLC Online Computer Library Center, Inc. Used with permission.

The survey results and evaluation of aggregation services are presented in five different
"problem spaces," reflecting the context in which they make their greatest contribution.

- **Points of Reference: Open Access and the Open Archives Initiative**
  This section describes the registries, directories, indexes, and general cross-archive search engines that serve as entry points for finding OA and OAI content and collections. This set of services provides the foundation for understanding the broad scope and sweep of digital repositories.

- **Links in the Scholarly Communication Value Chain**
  Stemming from the self-archiving of research output in e-print repositories, this section describes various subject domain aggregations along with affiliated discovery and citation analysis services. Connected together, they serve vital functions in the scholarly communication value chain: registration, certification, awareness, archiving, and rewarding (Van de Sompel et al. 2004).

- **Pathways to E-Learning in the Sciences and Beyond**
  This section describes the National Science Digital Library (NSDL) and several partner projects, alongside a complementary community of practice in e-learning (MERLOT). There services are increasingly anchored and sustained by discipline-based entities as they move from a collections-driven approach to an emphasis on pathways and community participation.

- **Joining Forces: Cultural Heritage and Humanities Scholarship**
  These aggregations leverage digital collections and content contributed by diverse cultural heritage institutions for use by multiple constituents, ranging from the interested public to the research community. Humanists are building the processes, tools, and structures to support their work in the realm of digital scholarship.

- **User Alchemy: Discover, Deliver, Divine**
  From academic search engines and Internet "mentors" to customized portals, these services increasingly put the user first, surrounding them with domain-specific tools and resources at the point of need.

PART II: CONTEXTS

2.0   Scholarly Information Environment 2006

Viewed from any angle, the topography of scholarly digital activity has changed markedly since the DLF released its original survey less than three years ago. At first glimpse, there are many reassuring landmarks—most of the core services discussed in the original report continue to tower over the landscape; however, on closer examination, the scene is anything but static. To highlight some of the changes:

- OAIster represents triple the number of repositories and nearly five times the number of records with links to full-content;
- CiteSeer, PubMed Central, and ROAR are OAI-compliant;
- the deposit of scholarly articles in Open Archives, such as PubMed Central, has become a matter of (inter)national policy debate;
- Arc, NSDL, DLESE, Cornucopia and Perseus (among others) have overhauled their technical architecture;
- Heritage Colorado now crosses state borders, broadening its outreach to become Heritage West;
- the Resource Discovery Network will soon re-emerge as "Intute," an "authoritative mentor of the Internet for educators and researchers";
- the new CIC Metadata Portal and DLF MODS Portal have opened A9.com access to facilitate simultaneous searching with Amazon, Wikipedia, RedLightGreen, The British Library catalog, and other OpenSearch services; and
- recent initiatives such as NINES, SouthComb, and DLF Aquifer are developing scholarly tools and services in support of humanities cyberinfrastructure.

When attempting to understand the scale and pace of change, three phenomena stand out:

1. the articulation of cyberinfrastructures or e-frameworks to support domain services in the sciences, social sciences, and humanities;
2. the international ascendance of the Open Access (OA) movement fueling the growth of OA repositories and journals; and
3. the "amazoogle effect" (Dempsey 2004) that places the user front and center.

The impact of these phenomena is far-reaching, controversial, and complicated. Each represents multiple communities of practice and is the subject of innumerable articles, reports, blogs, and conferences. An in-depth review of these factors is beyond the scope of this report, but they are discussed below in relation to the services under review in this report to provide a context to understanding key factors precipitating change.

2.1 Cyberinfrastructure (CI) Articulation

When the DLF issued the previous survey in 2003, digital repositories were only starting to formulate their contributions to cyberinfrastructure (CI) in the sciences and engineering; the corresponding reports on CI in the humanities and social sciences had yet to appear. In October 2004, a symposium jointly sponsored by the Association of Research Libraries and the Coalition of Networked Information launched an important dialogue across stakeholder groups about the function and contributions of libraries to e-research and cyberinfrastructure (Goldenberg-Hart 2004). Referring to transformative changes in scholarly practice, CNI's Executive Director Clifford Lynch advised libraries that "a failure to put into place effective new support structures in response to these changes would pose tremendous risk to the enterprise of research and scholarship. The role of libraries, he argued, will shift from primarily acquiring published scholarship to a broader role of managing scholarship in collaboration with the researchers that develop and draw upon it" (Ibid).[3]

With the creation of the Cyberinfrastructure Council and establishment of the Office of Cyberinfrastructure in 2005, the National Science Foundation (NSF) has put into place a management structure to oversee its growing investment in effective CI development and deployment. The CI-Team represents a "cross-cutting," NSF-wide activity in which all Directorates participate, including the Directorate for Undergraduate Education in the Division of Education and Human Resources (EHR) that oversees funding for the National Science Digital Library (NSDL), and the Directorate for Geosciences, that funds the Digital Library for Earth System Education (DLESE). NSDL and DLESE (both described later in this report) serve as bridges between the e-learning and e-research communities in cyberinfrastructure development. Especially noteworthy is DLESE's partnership with GEONgrid, a network building cyberinfrastructure for the geosciences, relying heavily on geoinformatics.

The cyberinfrastructure phenomenon has galvanized scientific communities into action, fueling the transformation of disciplines and ushering in new research methodologies.[4] The CI-Team's Web site (http://www.nsf.gov/crssprgm/ci-team/) provides links to relevant reports and projects, including discipline-specific endeavors such as GEONgrid.[5] It also identifies major CI reports and projects relevant to the humanities and social sciences. Several new projects described in this report reflect efforts to build

---

[3] Purdue University provides an excellent example of how librarians are actively contributing to e-research initiatives. Refer to the CNI Spring 2006 Task Force report by Mullins and Brandt available from http://www.cni.org/tfms/2006a.spring/abstracts/PB-mullins-building.html.

[4] It is beyond the scope of this report to explore the international dimensions of cyberinfrastructure but as one example refer to the ChinaGrid Overview (Jin, [2004]).

[5] See also the National Institute of Health's large-scale, data intensive project, caBIG: cancer BioInformatics Grid, https://cabig.nci.nih.gov/.

cyberinfrastructure support in the humanities. Cornucopia and the IMLS (Institute of Museum and Library Services) Digital Collections & Content exemplify the contributions of libraries, archives and museums while DLF Aquifer, NINES (A Networked Interface for 19th-Century Electronic Scholarship), and SouthComb reflect efforts to support digital scholarship in the humanities.

Released in January 2005, NSF's Cyberinfrastructure Vision For 21st Century Discovery, version 5.0, calls for the development of strategic plans for four key CI components:

- High Performance Computing;
- Data, Data Analysis, and Visualization;
- Collaboratories, Observatories, and Virtual Organizations; and
- Education and Workforce Development.

While these strategic efforts are associated most closely with scientific endeavors, they will also address many of the core values and issues evident throughout the DLF report, for example, achieving effective wide-scale interoperability; managing large-scale, heterogeneous resources; creating context-sensitive user applications; and preserving intellectual assets.

It comes as no surprise to find that capturing the research output from digital repositories such as arXiv, NTRS, and PubMed Central as well as in institutional repositories like the CERN Document Server (CDS) aggregation, is a key element in the design of complex, collaborative digital library infrastructures. NSDL aggregates metadata from these and other innovative services such as Reciprocal Net (http://www.reciprocalnet.org/) which harvests crystallographic data from 18 partner institutions and creates a searchable database. The JISC-funded eBank project, which is part of the UK's Semantic Grid Programme, demonstrates how to link "research data with other derived information" by harvesting from both e-print and e-data archives. Utilizing the GNU Eprints software, the project will link crystallographic data from the Combechem project with Intute's (formerly RDN) PSIgate Physical Sciences Information Gateway (http://www.ukoln.ac.uk/projects/ebank-uk/).[6]

Another pan-European initiative serves as an excellent example of transnational collaboration in building a digital research infrastructure. DRIVER: Digital Repository Infrastructure Vision for European Research is a consortium of nine European universities and research agencies that is developing a "test-bed for integrating existing national, regional and/or thematic repositories in order to create a production-quality European infrastructure" (Lossau 2006). DRIVER has four major activities: (1) content and organization provision through the aggregation of an initial set of 50 repositories;

---

[6] eBank project description from search at PSIgate at http://www.psigate.ac.uk/newsite/. Refer to Lyon and Coles 2004 for more details and graphic images of the architecture.

(2) implementation of an open, distributed, service-oriented repository infrastructure middleware; (3) focused studies; and (4) raising awareness. DRIVER expects to start in July 2006, release in June 2007, and become a production service in April 2008. The figure below illustrates the initial conceptualization of DRIVER's infrastructure.

Figure 03: Digital Repository Infrastructure Vision for European Research
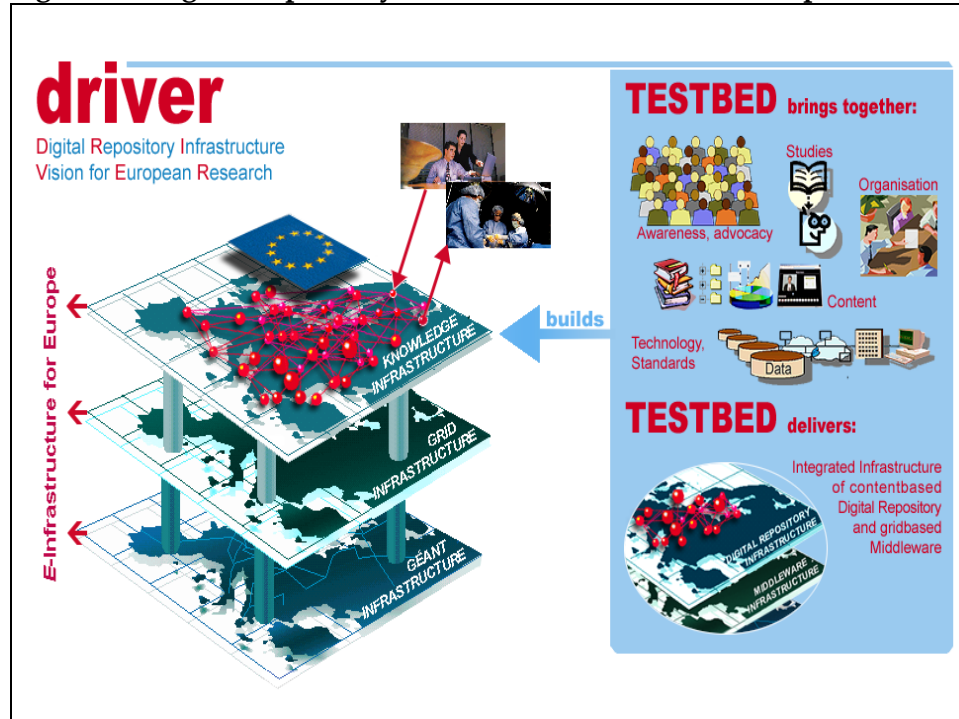
Source: Lossau 2006. Used with author's permission.

2.1.1 Convergence Across Higher Education Service Domains

Blinco and McLean's "cosmic view" is perhaps the best depiction of key variables that influence the digital repository workspace. Their conceptualization identifies five major service domains in higher education: E-Research, Scholarly Information, E-Learning, Web Publishing, and Administrative Computing. Designed as a "Wheel of Fortune" set in motion, the concentric circles rotate, illustrating the multiple ways that services may realign. It aptly illustrates the multiple contexts in which digital repositories operate, highlighting the need to develop flexible and coherent frameworks that manage these permutations.

Figure 04: 'Cosmic' View of Repositories Space (Wheel of Fortune)

A "Cosmic" View of the Repositories Space
(Wheel of Fortune)

© Kerry Blinco & Neil McLean, 2004
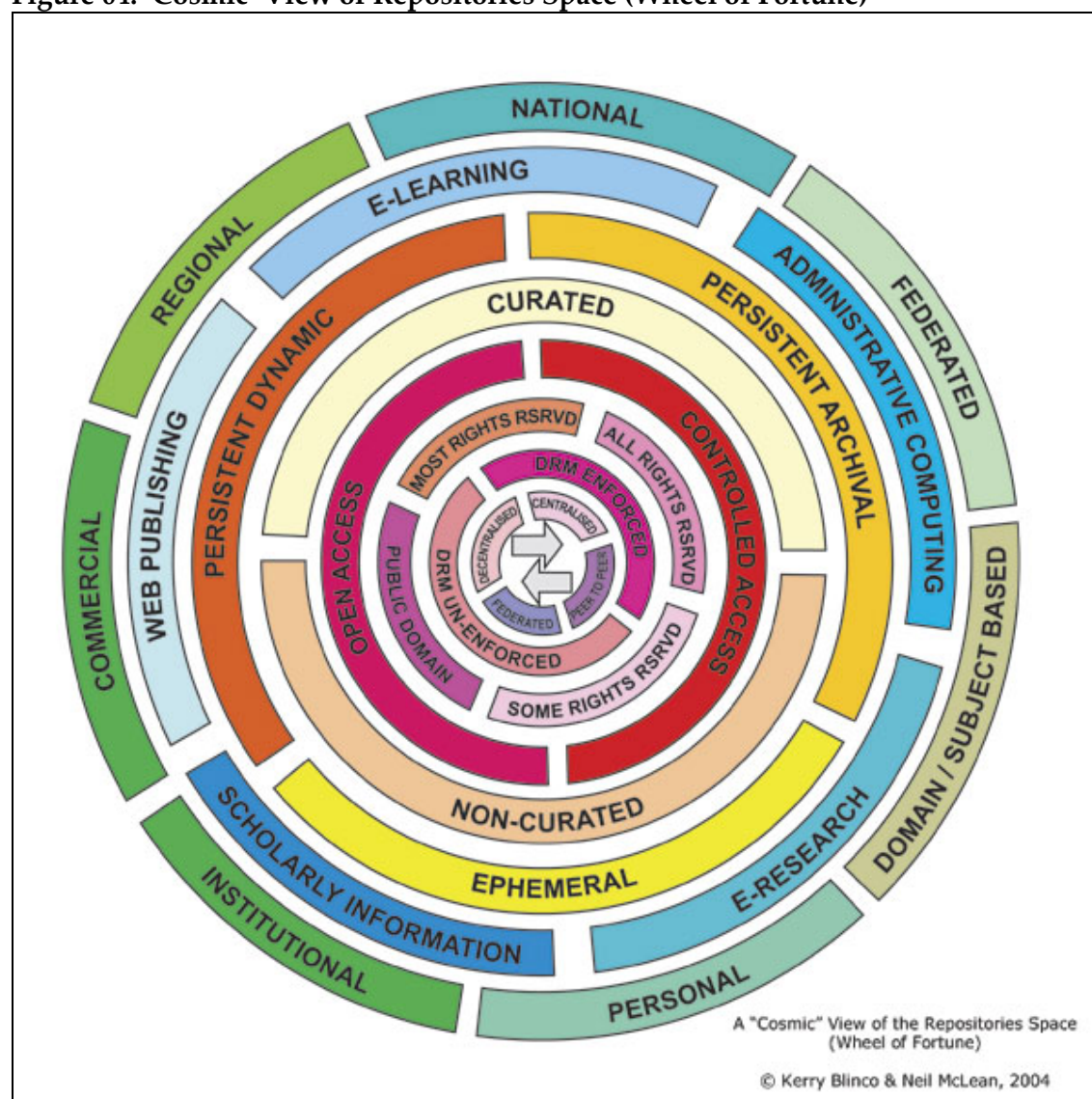
Reprinted with permission of the authors. Also available from
http://www.rubric.edu.au/extrafiles/wheel/.

The e-Framework for Education and Research (http://www.e-framework.org/), an initiative led by the UK's Joint Information Systems Committee (JISC) and Australia's Department of Education, Science and Technology (DEST), aptly illustrates this new information environment. It aims "to produce an evolving and sustainable, open standards based, service-oriented technical framework to support the education and research communities." The partnership's guiding principles espouse many of the values identified in the 2003 and 2006 DLF surveys, while situating them into a coherent structure:

1. The adoption of a service oriented approach to system and process integration.
2. The development, promotion and adoption of Open Standards.
3. Community involvement in the development of the e-Framework.
4. Open collaborative development activities.
5. Flexible and incremental deployment of the e-Framework
   (Source: http://www.e-framework.org/about/).

Taken together, these principles are international, collaborative, standards-based, platform independent, promote excellence through agreement on best practices, avoid high-level service duplication, encourage innovation, engage user communities, support both open source and proprietary implementations, and, are mindful of business practices.

The initiative draws on the Australian e-Learning Framework (http://www.elframework.org/) and JISC's Information Environment (JISC IE) architecture to facilitate interoperability across education and research domains. McLean (2004) charts out the "evolving e-learning framework" and identifies a layer of common services, which he then places into a broader context by mapping them across different domains. This practical exercise demonstrates how a service-oriented approach to frameworks and architectures helps to identify common services, and could lead to shared technical development across multiple domains. Clicking on components in the framework links to relevant projects, specifications, and discussion forum comments.

The JISC IE technical architecture "specifies a set of standards and protocols that support the development and delivery of an integrated set of networked services that allow the end-user to **discover, access, use** and **publish** digital and physical resources as part of their learning and research activities."

Figure 05: Screenshot of E-Learning Framework

Source: http://www.elframework.org/framework/ (May 2006)
 (http://www.ukoln.ac.uk/distributed-systems/jisc-ie/arch/).
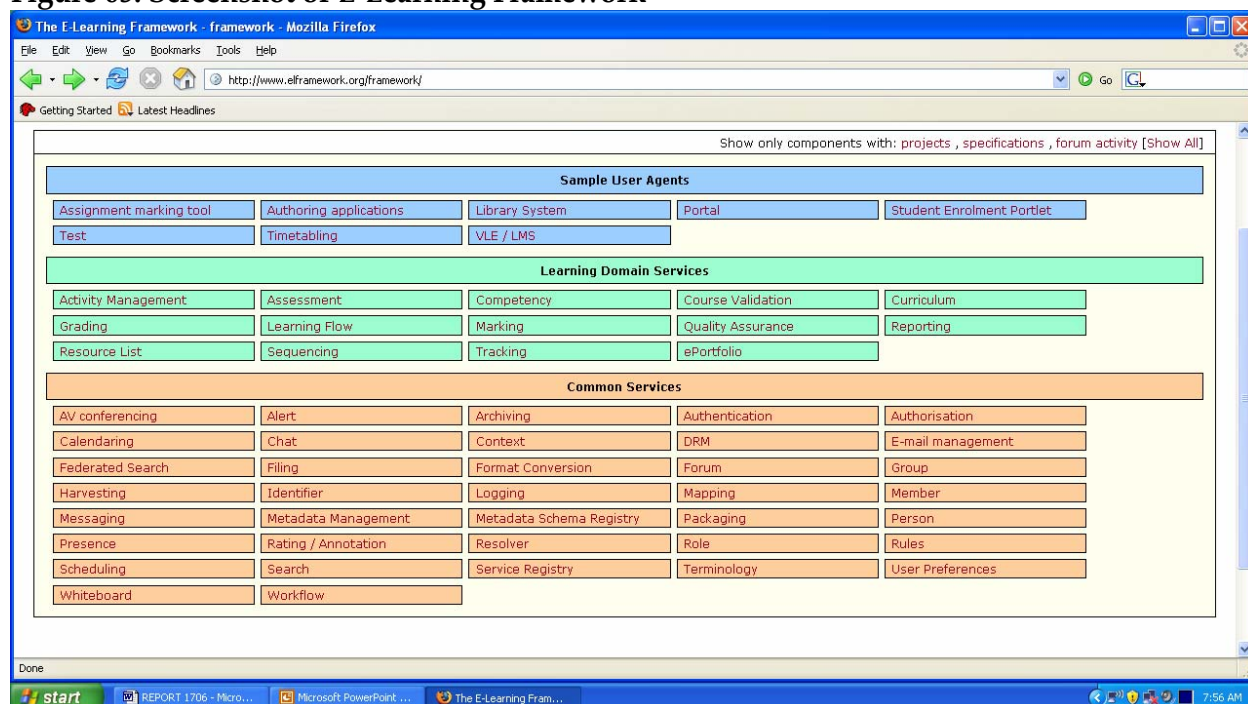
Figure 06: Screenshot of the JISC Information Environment Architecture
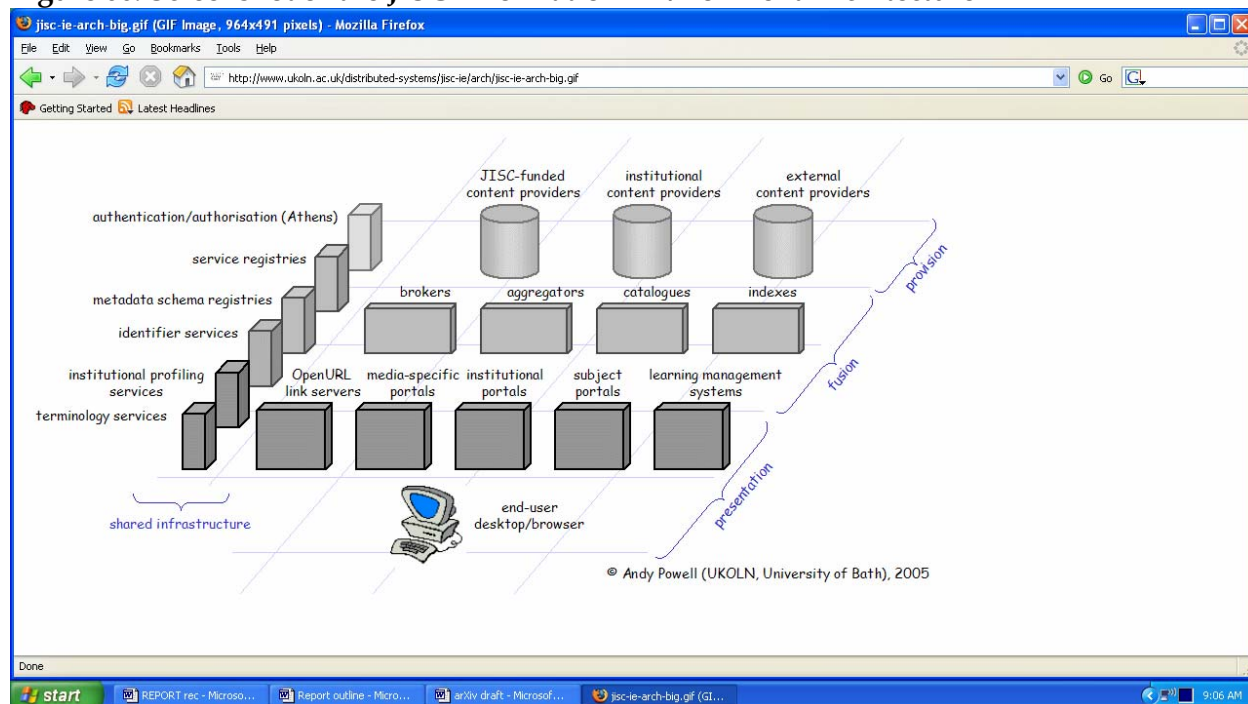
Source: http://www.ukoln.ac.uk/distributed-systems/jisc-ie/arch/jisc-ie-arch-big.gif © Andy
Powell (UKOLN, University of Bath), 2005.

The user environment features a shared infrastructure of common services running across the presentation, fusion, and provision layers. JISC's IE service framework supports such activities as: [7]

- Integration of local and remote information resources with a variety of 'discovery' services . . . allowing students, lecturers and researchers to find quality assured resources from a wide range of content providers including commercial content providers and those within the higher and further education community and elsewhere.
- Seamless linking from 'discovery' services to appropriate 'delivery' services.
- Integration of information resources and learning object repositories with Virtual Learning Environments….
- Open access to e-print archives and other systems for managing the intellectual output of institutions. (Powell 2005).

2.1.2 Discipline-based Landscape Analysis[8]

If the "Wheel of Fortune" scopes out a high-level e-framework and corresponding service-oriented architecture (SOA), then PerX, a pilot UK project in engineering, offers a model for producing a discipline-specific landscape analysis and corresponding specialized cross-archive search system. The PerX project compiled an inventory of existing and potential engineering repository sources, using Heery and Anderson's (2005) basic typology that first, distinguishes repositories with content from metadata-only repositories and secondly, classifies them by content type, coverage, primary functionality and target user group (http://www.icbl.hw.ac.uk/perx/sourceslisting.htm#broadlandscape).

Table 02: Repository Typology, modified and abbreviated version of Heery and Anderson 2005

Via Content Type	Via Primary Functionality
- Research Data - Research Outputs - e-Theses - Learning Materials - Multimedia - Assessment Materials	- Subject access to resources - Enhanced access to resources - Preservation of digital resources - New Modes of dissemination/Publication - Institutional Asset Management

---

[7] For information about the Digital Library Federation's parallel conceptualization refer to the "DLF Service Framework for Digital Libraries," a progress report for the DLF Steering Committee, prepared by Lorcan Dempsey and Brian Lavoie, May 17, 2005. Available from http://www.diglib.org/architectures/serviceframe/dlfserviceframe1.htm

[8] For an extensive survey about disciplinary differences refer to the "JISC Disciplinary Differences Report" (Sparks 2005), available from JISC's Committee for the Information Environment, Scholarly Communications Group, http://www.jisc.ac.uk/index.cfm?name=schol_comms_reports.

• Corporate Records	• Sharing and Reuse of Resources
Via Coverage	**Via Target User Group**
• Personal/Informal	• Learners
• Journal	• Teachers
• Institutional/Departmental	• Researchers
• Inter-Institutional	
• National	
• Geospatial	

Source: http://www.icbl.hw.ac.uk/perx/sourceslisting.htm#broadlandscape

PerX then used the source listing as the basis for analyzing the position of repositories by content type (http://www.icbl.hw.ac.uk/perx/analysis.htm). The Table below gives a synopsis of the status of sources that support engineering preprints/postprints and technical reports.

Table 03: Extract of PerX by Repository Content Type with Synopsis of Position

Research Outputs 1: Preprints/Postprints
Repositories:
• Wide scale adoption of Institutional Repositories seems likely following on from impetus arising from the UK Select Committee on Science and Technology Report and subsequent activities.
• There are a growing number of institutional repositories in the UK. Adoption is being encouraged through various initiatives such as SHERPA and the Digital Repositories Programme.
• As yet there is no established National infrastructure for coordination of UK HE research output. Swan et al (2005) discusses a model for eprint content in the UK and the ARROW project provides an example of a national resource discovery service for research outputs in Australia.
• Most institutional repositories contain multidisciplinary material. There are often no means available for a subject based service to select subsets of such multidisciplinary collections based on subject coverage. Clearly OAI-PMH Sets could be used but in practice few repositories provide subject based sets.
Metadata Repositories:
• No general purpose engineering Preprints/Postprints metadata repositories have been identified
• A number of more specialised metadata repositories exist (e.g. ASEE Conference Proceedings).
Research Outputs 2: Technical Reports
Repositories
• Gap area
• A number of substantial engineering technical report repositories are in existence, which cover technical reports published in the USA (e.g. NASA, NACA).
• There are few equivalents for the UK, one exception being reports available via the CCLRC ePublication Archive. On the whole, bibliographic control of technical report series tends to be complex, they are often poorly catalogued, and scattered across the academic, government and commercial sectors. This makes technical reports difficult to identify, locate and access.

> - The MAGIC project investigated issues surrounding access to the technical reports, and produced the METReS demonstrator service, which was in essence a prototype UK National Technical Reports Catalogue. The MAGIC project ended in Oct 2002 and the METReS service is no longer supported and is available only for archival purposes.
> **Metadata Repositories**
> - A number of technical report metadata repositories are in existence. Coverage of US technical reports is good (NTIS, STINET). Other metadata repositories in this area tend to include various types of 'grey' literature as well as technical reports (Energy Citations Database, GrayLit Network, etc)

Source: http://www.icbl.hw.ac.uk/perx/analysis.htm Engineering Digital Repositories Landscape Analysis, and Implications for PerX, Version 1.0 (MacLeod and Moffat 2005).

The landscape analysis further identifies the information and communication needs of engineers and describes the complexity of the published engineering information landscape. Next, a pilot subject-based cross-archive search system was created, featuring the different repositories identified through the landscape analysis (http://www.engineering.ac.uk/). Other project deliverables include development of "advocacy materials," such as the excellent Web accessible document—"Marketing with Metadata: How Metadata Can Increase Exposure and Visibility of Online Content," and embedding PerX into virtual learning environments (http://www.icbl.hw.ac.uk/perx/deliverables.htm) .

PerX has developed a model for analyzing needs and mapping them against resources that is readily adaptable to other disciplines.

2.2 Open Access Ascendant: Growth of OAI-compliant Repositories

The open access movement:
Putting peer-reviewed scientific and scholarly literature on the Internet.
Making it available free of charge and free of most copyright and licensing restrictions. Removing the barriers to serious research. Peter Suber, Open Access News.[9]

In the short span of time since "A Survey of Digital Library Aggregation Services" appeared, the open access movement has gained international momentum and engendered a multitude of commitments from major funding agencies, intergovernmental organizations, private and public foundations, university and library

---

[9] Those new to the concept of Open Access should refer to Suber's "Open Access Overview" http://www.earlham.edu/~peters/fos/overview.htm. Bailey (2006) examines three core definitions of open access; notes key OA statements; discusses self-archiving strategies and practices. For a more extensive consideration, refer to *The Access Principle: The Case of Open Access to Research and Scholarship* (Willinsky 2006).

consortia, and publishers. Glancing back to September 2003, it is instructive to recall how much the situation has changed: [10]

- the Berlin Declaration on Open Access to Knowledge in the Science and Humanities did not exist;
- the UN World Summit on the Information Society had not convened nor approved a Declaration of Principles and Plan of Action endorsing open access to scientific information;
- the U.S. National Institutes of Health and United Kingdom's Wellcome Trust did not have public-access policies;
- the Alliance for Taxpayer Access and the Open Content Alliance had not yet formed;
- the Public Library of Science had not launched its first OA journal, PLoS Biology;
- Elsevier, Springer, and SAGE did not permit authors to archive post-prints;
- PubMed Central's repositories of journal articles were not yet OAI-compliant;
- the Directory of Open Access Journals did not offer article-level access; and
- Cornyn and Lieberman had not introduced the Federal Research Public Access Act to the U.S. Senate, mandating OA to most federally funded research.

The Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities of October 2003, the third major public statement in support of open access, clinched the transition of open access (OA) principles from periphery to mainstream in the world of policymakers and digital library developers alike. While the three foundational declarations, (Budapest, Bethesda, Berlin), vary in nuance and domain, they share two complementary strategies for promulgating OA, namely, self-archiving (depositing scholarly articles in open electronic archives) and open access journals (either through the transition of existing, or creation of new journals).

With 164 organizational national and international signatories as of May 2006, the Berlin Declaration continues to evolve through annual conferences and by following the "Roadmap" devised in 2004 to implement Open Access. This influential 10-step plan provides guidelines for raising awareness, developing organizational policies, creating a sustainable infrastructure and legal framework, supporting OA journals, and securing long-term organizational commitments.

ROARMAP, the Registry of Open Access Repository Material Archiving Policies, (described more fully in section 4.1.4), maintained by EPrints.org at the University of Southampton, tracks policies and mandates worldwide, as recommended by the Berlin Declaration (http://www.eprints.org/openaccess/policysignup/).

---

[10] Sampling compiled from Peter Suber's "Timeline of the Open Access Movement" (last revised May 7, 2006). Available from http://www.earlham.edu/~peters/fos/timeline.htm.

The case for institutional repositories (IR) as a "critical component in reforming the system of scholarly communication" (Crow 2002) and as "essential infrastructure for scholarship in the Digital Age" (Lynch 2003) is closely aligned, if not frequently indistinguishable from the OA movement. Arguments for institutional repositories as a vehicle for opening up access to research are available in:

- Suber's "Open Access Overview"
  http://www.earlham.edu/~peters/fos/overview.htm;
- JISC's Questions and answers about opening up research results
  http://www.jisc.ac.uk/index.cfm?name=issue_qaopen;
- EPrints.org: Self-archiving FAQ. http://www.eprints.org/openaccess/self-faq/;
  and
- Steve Hitchcock's (Open Citation Project, EPrints.org) online bibliography, "The Effect of Open Access and Downloads ('Hits') on Citation Impact," selectively annotates relevant studies http://opcit.eprints.org/oacitation-biblio.html.

As evident from the analysis of "OAI demographics" below, OA archives are operational in at least 46 countries. The "top 20" implementations, in terms of size, come from 11 different countries. In addition to OAI deployments in the U.S., U.K., Australia, and five European countries, Brazil, Japan, and Mexico have significant OAI deployments—and there is little doubt that India (Sreekumar 2006, Ghosh 2006) and China (Tansley 2006) will soon join this list. Many developing and transitioning countries view OA repositories as a launch pad capable of bringing indigenous research output into the international arena, increasing its visibility and impact, while also "building research capacity" (Chan et al. 2005). Concurrently, a growing number of high-profile international projects and networks are emerging, designed to cross the "digital divide" and deliver high-quality scientific literature equitably:

- International Network for the Availability of Scientific Publications
  http://www.inasp.info/.
- CODATA (International Council for Science : Committee on Data for Science and Technology): http://www.codata.org/message-from-president.html.
- HINARI: Health InterNetwork Access to Research Initiative (World Health Organization) http://www.who.int/hinari/en/.
- AGORA -- Access to Global Online Research in Agriculture (Food and Agriculture Organization): http://www.aginternetwork.org/en/.
- OARE -- Online Access to Research in the Environment (The United Nations Environment Programme, Yale University, WHO, FAO, Cornell University and several leading publishers): http://www.unep.org/library/OARE_project.asp.

Two surveys carried out in conjunction with conferences held in 2005 and 2006 provide a wealth of information about IR and ETD (electronic theses and dissertations) deployments in Europe and the United States. In May 2005, representatives from 13

countries came together to discuss IRs under the co-sponsorship of the Coalition for Networked Information (CNI), the SURF Foundation of the Netherlands, and UK JISC. Two ensuing *D-Lib Magazine* articles discuss "Institutional Repository Deployments in the United States as of Early 2005" (Lynch and Lippincott 2005) and "Academic Institutional Repositories: Deployment Status in 13 Nations as of Mid 2005" (van Westrienen and Lynch 2005). While these articles provide useful overviews and preliminary data, the corresponding questionnaires also contain a wealth of information about country-specific deployments (van Westrienen 2005). Similarly, the JISC-SURF-CURL (Consortium of Research Libraries in the British Isles) sponsored "International Workshop on E-Theses" (Jacobs 2006a) has a corresponding compilation of country-specific responses about the status of ETD programs in 11 European countries. The DLF report draws on a handful of these country-specific examples, especially in the section pertaining to "Links in the Scholarly Communication Value Chain."

2.2.1 Enabling OA Technology Platforms

Institutional repositories constitute one service model by which to achieve the Open Access agenda, and enabling applications such as DSpace and EPrints.org, are the open-source software technology platforms to realize this goal.[11] MacKenzie Smith notes that DSpace distinguishes itself from typical open source software projects in several ways:

- Its users are organizations, not individuals;
- DSpace is an entire application, not a tool;
- DSpace is an end-user application, not middleware or productivity tool; and
- Features and functions decided by domain experts, not programmers. (Smith 2006)

DSpace (MIT), EPrints.org (University of Southampton), CDSware (CERN, Switzerland), Achimède (University of Laval), OPUS (University of Stuttgart), and Fedora (Cornell and University of Virginia), are international communities of practice, spawning innovation to meet the service needs of their user communities.[12] In effect, these systems are simultaneously service models—opening access to research information through self-archiving—and technology platforms—capturing, diffusing and archiving intellectual output; the two functions inextricably bound together. EPrints.org, for example, offers its users a menu of services including advising on policy matters; training; assisting with advocacy and IR promotion; importing legacy archives;

---

[11] Although IR adoptions are emphasized here, it is worth noting that most of these software systems are used to manage a range of services, for example learning object repositories, e-theses, electronic records management, e-publishing, and digital preservation. See EPrints.org examples of different types of deployments: http://www.eprints.org/software/examples/.

[12] The Appendices to *Institutional Repositories* (Jones et al. 2006) have basic descriptions of each of these software projects.

customizing; hosting and maintaining the repository; and providing ongoing technical support.[13] A JISC-funded project, IRRA (Institutional Repositories and Research Assessment) is creating mechanisms to mesh DSpace and EPrints.org workflows with the UK's Research Assessment Exercise (RAE), which will move to a metrics-based methodology after 2008 (UK. HM Treasury 2006). RAE is a national assessment of the quality of research that informs the distribution of public research funds in the UK (http://www.rae.ac.uk/). Among other efforts, the DSpace and EPrints.org communities are mapping input options (item types) to match RAE output types (e.g., edited book, journal article, conference contribution). [14]

Open source software developed by the Public Knowledge Project (PKP), led by the University of British Columbia and Simon Fraser University, facilitates the uptake of OA journals (via OJS, the Open Journal Systems) and conferences proceedings (via OCS, the Open Conference Systems). OJS is a "journal management and publishing system" that "assists with every stage of the refereed publishing process, from submissions through to online publication and indexing" (http://pkp.sfu.ca/). With a worldwide user community representing more than 550 journals, OJS supports the African Journals Online project (more than 200 titles) and the Brazilian Institute of Science and Technology Information (more than 80 titles).

In addition to open source projects, a growing number of semi-proprietary and commercial vendors are offering IR services.[15] These technology platforms also typically have OAI interfaces, facilitating the exposure and harvesting of metadata. Bepress (Berkeley Electronic Press) licenses its repository software technology to ProQuest Information and Learning. Known as Digital Commons, ProQuest registers its client repositories (as of May 2006, 50 institutional and consortial customers) with OAIster and also manages data transfer to other third-party services and indexes such as Google and Yahoo!Search. The New England Law Library Repository (http://lsr.nellco.org/) uses Digital Commons/bepress software to aggregate research papers from the 25 member institutions in the NELLCO consortium. Similarly, COBRA: the Collection of Biostatistics Research Archive, is a bepress subject repository of prepublication biostatistical materials contributed by 13 institutions (http://www.biostatsresearch.com/). Bepress's award-winning "quasi-open access" aggregation of journals, IR contents and subject-based archives, ResearchNow (http://researchnow.bepress.com/), is discussed in section 4.2.11.

---

[13] See "Types of Service Offered" available from http://www.eprints.org/services/.

[14] The RAE submission guidelines for Research outputs are available from http://www.rae.ac.uk/datacoll/subs/.

[15] See DLF's "Tools document," prepared as part of the DLF OAI curriculum and training materials, discussed in section 3.1.1, for more examples of digital content management systems and vendors.

2.2.2 OAI Demographics 2006

It is possible to formulate a composite picture of OA deployments by examining data from several OAI and OA registries and databases. As of May 2006, the University of Illinois OAI-PMH Data Provider Registry lists nearly 1,050 active OAI-compliant repositories (data providers)—or five times the estimated number of deployments two and a half years ago. As explained more fully in section 4.1.1, the UIUC registry strives to be comprehensive and deploys a systematic multi-faceted approach (that goes beyond self-registration) to achieve the goal of completeness.

Meanwhile data from two other sources, ROAR (Registry of Open Access Repositories) and OAIster, (both described in section 4.1), affords an in-depth comparison of the geographic distribution and size of OA/OAI repositories. As of mid-March 2006, ROAR listed 640 repositories and OAIster, 597, representing 46 countries altogether.

Table 04: Comparison of ROAR and OAIster Data

	ROAR	OAIster
Number of Repositories	640	597
Number of Countries	40	42
Unique Countries (not represented in the other database)	Costa Rica, Israel, Namibia, Pakistan, Turkey, Singapore	Czech Republic, Lithuania, Poland, Venezuela, Serbia, West Indies
Top Three Countries	United States: 178 (27.8%) United Kingdom: 69 (10.8%) Germany: 60 (9.4%) *Top 3 countries constitute 48% of the repositories.	United States: 223 (37.4%) United Kingdom: 63 (10.5%) Germany: 58 (9.5%) *Top 3 countries constitute 57.4% of the repositories.
Countries with 10 or more implementations besides the US, UK, and Germany	In descending order: Brazil, Canada, France, Australia, Sweden, Italy, Netherlands, India, Spain, and Other. *Constitute 36.4% of total.	In descending order: Canada, France, Brazil, Australia, Netherlands, Italy, Sweden, Spain, and Multinational. *Constitute 28% of total.
Countries with only one implementation	9 countries	11 countries
Number of records per repository:	7,767 (average) 142 (median)	11,727 (average) 542 (median)
Number of repositories with fewer than 50 OAI records	114	106
Number of repositories with fewer than 10 OAI records	40	27

ROAR

- Open access implementations are operational in 40 countries.
- Of the 46 countries combined in ROAR and OAIster, six are unique to ROAR: Costa Rica, Israel, Namibia, Pakistan, Turkey, and Singapore.
- With 178 instantiations, the United States (27.8 percent) leads the list, followed by the United Kingdom with 69 (10.8 percent) and Germany with 60 (9.4 percent). These three countries constitute almost 50 percent of all implementations.
- Another 9 countries and one "other" category have ten or more implementations and comprise 36.4 percent of the total (in descending order): Brazil, Canada, France, Australia, Sweden, Italy, the Netherlands, India, Spain, and Other.
- Nine countries are represented by only one instantiation.
- The average number of records per repository (those represented in Celestial) is 7,767. The median number of OAI records among these 480 archives is 142. An estimated 114 repositories have fewer than 50 OAI records; 40 have fewer than ten OAI records.

OAIster

- OAI-compliant implementations are in operation in 42 countries worldwide.
- Of the 46 countries combined in ROAR and OAIster, six are unique to OAIster: Czech Republic, Lithuania, Poland, Venezuela, Serbia, and West Indies.
- With 223 instantiations, the United States (37.4 percent) leads the list, followed by the UK with 63 (10.5 percent) and Germany 58 (9.5 percent). These three countries constitute almost 60 percent of all implementations.
- Another eight countries and a "multi-national" category have ten or more implementations, comprising 28 percent of the total (in descending order): Canada, France, Brazil, Australia, Netherlands, Italy, Sweden, Spain, and Multinational.
- Eleven countries are represented by one instantiation.
- The average number of records per repository is 11,727 items. The median is 542 among the 597 repositories. An estimated 106 have fewer than 50 OAI records; 27 have fewer than 10 OAI records.

Additional OAIster data about Repositories in the United States

- OAIster harvests from repositories located in 38 different states. Twelve states have no records, unless they are represented among the ten repositories with multi-state services (accounting for some 240,000 records). States without any representation in OAIster include: Arkansas, Delaware, Idaho, Maine, Missouri, Montana, New Hampshire, North Dakota, South Dakota, Vermont, West Virginia, and Wyoming.
- In terms of the number of items in the 223 US repositories represented in OAIster:
    - 8 have > 100,000 records
    - 36 have > 10,000 records
    - 88 have < 500 records
    - 52 have < 100 records

Top 20 Repositories in ROAR and OAIster

While quality and quantity are not synonymous, critical mass is important and the sheer number of records in a repository indicates an area of concentrated activity. Again, a comparison of the top 20 archives in terms of record count in ROAR and OAIster reveals both commonalities and differences. When comparing ROAR and OAIster data, it is important to bear in mind key differences in their harvesting parameters and purpose. ROAR relies on self-registration and focuses on e-print archives, especially those that use GNU EPrints software. Meanwhile, OAIster harvests from OAI-compliant collections of all media types (including images and datasets) and includes some repositories that restrict use to licensed users (e.g., Institute of Physics journal articles). Secondly, ROAR harvests metadata records without requiring them to point to full-content digital objects (i.e. metadata only), whereas OAIster only harvests those records with full-content representation. Third, ROAR only has OAI record counts for the subset of its 640 archives harvested by Celestial (480 repositories). In contrast, OAIster has item counts for all 597 of its repositories. Fourth, there are differences in whether they harvest some or all of the available records from any given service. For example, OAIster recently began to harvest OSTI's technical reports directly rather than through the NSDL aggregation, thereby causing NSDL to tumble off OAIster's list of 20 largest repositories, and to propel OSTI onto it. Fifth, the actual harvests take place at different intervals so counts may lag behind at any given time. For example, RePEc (Research Papers in Economics) merged records from the American Economic Association's working paper collection into its aggregation in March 2006, causing it to shoot up in size from around 50,000 to more than 130,000 records. When this Top 20 snapshot comparison was undertaken, it is probable (given its low record count) that ROAR had not yet harvested from the RePEc enlarged collection. Finally, the record counts within either service include an undetermined number of duplicates.

The complete top 20 list is available in Appendix 4. The summary of findings follows:

- The top 20 ROAR archives have 36,000 or more OAI records with a combined total of more than 3 million, accounting for 81 percent of the total record count. CiteSeer is the largest (> 700,000 records) followed by PubMed Central with nearly 500,000 records.
- OAIster's top 20 archives, those with 79,000 or more items, account for 72 percent of OAIster's 7 million records. The Australian consortial image database, Picture Australia, is the largest with more than 800,000 items, followed by CiteSeer.
- With different top 20 record-count thresholds and distinct harvesting parameters, only six of ROAR's top 20 figure among OAIster's top list: CiteSeer, PubMed Central, arXiv, Library of Congress Digitized Historical Collections, the National Institute of Informatics (Japan), and RePEc. Once OAIster achieves its top 20 threshold, its record count for a number of archives closely parallels other ROAR top

20 repositories: DIALNET, SciELO, HAL, Demetrius Australia National University, and Gallica.

- The combined top 20 lists represent 33 different archives in eleven different countries: Australia, Brazil, France, Germany, Japan, Mexico, Netherlands, Spain, Switzerland, United Kingdom, and United States.
- Half of them focus predominantly in the scientific, biomedical and technical disciplines and represent aggregations of e-prints, technical reports, journal articles, datasets, mathematical formulas, and citation analysis and alerting services. They are hosted by a variety of entities ranging from international scientific laboratories and universities to national funding agencies and learned societies.
- About a quarter, spawned by national libraries and universities, represent digital aggregations of images, maps, manuscripts, sound recordings, and photographs drawn primarily from historical and special collections such as those of the Library of Congress, National Library of Australia, and the University of Southern California.
- The remaining 25 percent represent a mix of institutional repositories (including three from Dutch universities), electronic theses and dissertations, social science e-prints and working papers, and U.S. state government reports.

In conclusion, ROAR and OAIster are useful tools for analyzing patterns in OAI adoption and growth. As discussed more fully in section 4.1.4, ROAR tracks the growth of repositories at the individual and composite level. It also makes statistics available by country, archive type and software in use. OAIster sends historical data to the UIUC registry on a monthly basis, making it possible to view growth data by repository; however, it is not accessible in a very user-friendly form. OAIster also provides underlying data upon special request. As the above exercise demonstrates, it would be of great benefit to the research community if the UIUC registry and/or OAIster made their database management information Web accessible in user-friendly (downloadable, malleable) form. The data begs for wider exploitation and analysis to gain a better understanding of OAI implementations throughout the world.

2.3   "The 'Amazoogle' Effect"

In 2003, "Amazoogle" was not in our vocabulary (Dempsey 2004) or integral to the scholarly *Zeitgeist*. Google Scholar (http://scholar.google.com/), Google Books (http://books.google.com/), and the Google 5 (http://books.google.com/googlebooks/library.html) had not launched. Amazon's A9 (http://www.a9.com/), Flickr (http://www.flickr.com/about/), del.icio.us (http://del.icio.us/about/), and Connotea (http://www.connotea.org/faq) were fantasies. WorldCat was not "open" (http://www.oclc.org/worldcat/open/default.htm), nor accepting user-contributed content (http://www.oclc.org/productworks/wcwiki.htm).

Recognizing the undeniable magnetism of such enterprises, Dempsey (2005) urged the library community to re-consider the nature of library services in the context of these "major web presences which have become the first—and sometimes last—resort of research for many of our users." He evaluated the typical library user's experience against a set of common attributes inherent in popular Web services:

1. A comprehensive discovery experience.
2. Predictable, often immediate, fulfillment.
3. Open to consumers.
4. Open to intermediate consumers.
5. A co-created experience.

In contrast to mainstream Web services, library "systems have low gravitational pull," Dempsey concluded, because "they do not put the user in control, they do not adapt reflexively based on user behavior," and "they do not participate fully in the network experience of their users."

2.3.1 What Recent User Studies Reveal

OCLC's extensive international survey of information consumers portrays a sobering reality for libraries: *only one percent of respondents begin an information search on a library Web site—84 percent use search engines first*; moreover, *90 percent of respondents are satisfied with their most recent search for information using a search engine.* Rather than speed, *the quality and quantity of information returned in the search process* is of primary importance. *Search engines fit the information consumer's lifestyle better than physical or online libraries.* While college students report the *highest rate of library use and broadest use of library resources. both physical and electronic, only 10 percent indicated that their library's collection fulfilled their information needs after accessing the library Web site from a search engine.* (Excerpted from De Rosa et al. 2005, 6-2, 6-3).

A wide-ranging survey of faculty attitudes and behaviors in using digital resources in undergraduate humanities and social sciences education found that "the most cited reasons for *not* using digital resources was that they simply do not mesh with faculty members pedagogies" (Harley et al. 2006, 180). The authors exhort:

> We should not expect faculty, who we can assume know more about teaching their subject than non-specialists, to shoehorn their approaches into a technical developer's ideas of what is valuable or the correct pedagogical approach. Tools and resources need to be developed to support what faculty do (Ibid).

In terms of integrating resources into learning management systems, the findings mirror those of the 2003 DLF survey of aggregations and DLF's Scholarly Advisory Panels (DLF 2004 and 2005):

- The difficulty, if not current impossibility, of re-aggregating objects that are bundled and "locked" into fixed, often proprietary resources.
- Managing and interpreting digital rights, which may include pulling data from one resource from one resource for integration into another.
- The unevenness of interface usability and aesthetics. (In some disciplines, such as art history, faculty may care a lot about resolution quality. Yet in other disciplines, faculty may create "hodgepodge" resources, often not caring about varying resolution quality from one record to the next.)
- The growing demand from users for granularity (e.g., the ability to search and find the one particular image or piece of text they need within an entire resource).
- The issue of knowing about and finding digital objects. Simply put, many faculty have no idea about the existence of local and non-local resources, especially licensed ones, which may be available to them (Ibid).

When reviewing existing OAI services (such as OAIster) in 2005, DLF's Scholarly Advisory Panel also reiterated many of the shortcomings identified in the 2003 DLF survey. They underscored the importance of understanding the context of retrieved results and the need for authoritative collection and item-level descriptions. If not all records carried full descriptive metadata, they wondered: what proportion of the database is queried when search delimiters such as, subject, date or author, are invoked?

In viewing results, scholars frequently could not distinguish items from collections, nor could they easily identify the collection to which an item belonged. They were critical of the search functionality, which gave precedence to institutions—over collections and subjects—when retrieving or sorting results. As a default, they preferred a basic search box, with a layered option for advanced search queries. They also wanted brief record results returned first or the option of letting users specify whether they wanted brief or full-records returned.

Turning to more advanced functionality, scholars hoped that OAI services would support:

- browsing by subject and collection as well as searching
- clustering and categorization by subject for browsing/searching ("search within" features)
- gathering records (e.g., in a personal book bag) to create virtual collections for use in the classroom
- viewing and downloading the metadata alongside the data (to support export into citation software or for annotating)

- thumbnail grabber or text grabber services to enrich metadata records with evidence of what they describe
- alerting services targeted to their particular interests
- extending searches to other indexing services (e.g., to Google) or to find "more items like this."

The DLF MODS Portal, described more fully in section 4.1.8, is an early attempt to implement some of the features and functionality considered most desirable by this representative group of humanists.

2.3.2 Creating User-focused Services

In an effort to meet user expectations brought to light in these and similar studies, the University of California Libraries set out to re-examine the way in which it delivered bibliographic services to users. The ensuing report identifies a desirable set of enhanced search and retrieval capabilities to apply throughout the University of California library system. The UC list parallels the types of enhancements aspired to, or implemented by the services under review here:

1. Provide users with direct access to item
2. Provide recommender features
3. Support customization/personalization
4. Offer alternative actions for failed or suspect searches
5. Offer better navigation of large sets of search results
6. Deliver bibliographic services where the users are
7. Provide better searching for non-Roman materials (University of California Libraries Bibliographic Services Task Force 2005)[16]

In tandem, UC's Melvyl Recommender Project is exploring how to incorporate five key components into its next generation online public access catalog: relevance ranking, recommending, auto-correction, text-based discovery system, and new user interface strategies such as faceted browsing for subject groupings and bibliographic record grouping (http://www.cdlib.org/inside/projects/melvyl_recommender/).

---

[16] Full report available from http://libraries.universityofcalifornia.edu/sopag/BSTF/Final.pdf. Executive summary from CNI presentation available from http://www.cni.org/tfms/2006a.spring/abstracts/PB-whitney-melvyl.html.

Figure 07: Faceted Browse, Melvyl Recommender Project, California Digital Library

Source: (Whitney and Brantley 2006) http://www.cni.org/tfms/2006a.spring/abstracts/PB-whitney-melvyl.html.

The California Digital Library's Metasearch Initiative (described in section 4.5.4) illustrates an infrastructure that will bring together the panoply of locally held, centrally harvested, and externally located resources under a common framework and present them to the user in the context of their needs through different portals (see figure 07 and figure 45). At a national level, the NISO Metasearch Initiative brings together the various stakeholders—content providers, software providers and implementing libraries—in an effort to develop standards and best practices that will enable cohesive search and retrieval across disparate resources and platforms (Hodgson, Pace and Walker 2006).

Many, one is tempted to say all, of the services under review in this report demonstrate ways in which user-driven services are increasingly integrated into the overall scholarly information environment. Examples abound, ranging from direct access to full content (OAIster) and disciplinary pathways (NSDL) to peer-review of resources (BEN, DLESE, MERLOT, NINES); expert commentary (CiteSeer, NINES, NSDL, Perseus, SouthComb); and recommender systems (CDS, Scirus, Perseus). Moreover, as is the case with CDL, the new architectures deployed or under development at such services as NSDL, NEEDS, NINES, and Intute aim to enable a customizable, context-sensitive user experience. It is too early to know how these (mostly) nascent efforts will fare or to determine if scholars will find the resources sufficiently valuable and the tools easy

enough to use to integrate them into their crowded daily routines. Of course, personalization, visualization, and social networking tools are only valuable if the content and digital resources themselves are high quality and compatible with scholars' instructional or research methods.

PART III: CONTRIBUTIONS

3.0 Next Generation OAI

As hoped, OAI's flexibility and relatively low common denominator of required elements has helped to foster adoption by a wide range of domains and institutions. The specification for OAI Static Repositories and Gateways, released in October 2004, fosters growth among smaller, resource-challenged repositories in cases where OAI implementation would otherwise have proven beyond their capacity (Habing 2005). The specification provides a simple means for small collections (fewer than 5,000 records) that do not change frequently (less often than monthly) to expose metadata in a single XML file (10-20 MB) for harvesting through intermediation of an OAI Static Gateway.[17] In February 2006, JISC (UK) announced the STARGATE project, Static Repository Gateway Toolkit: Enabling small publishers to participate in OAI-PMH-based services. According to the press release:

> The project is implementing a series of static repositories of publisher metadata, and will demonstrate the interoperability of the exposed metadata through harvesting and cross-searching via a static repository gateway, and conduct a critical evaluation of the static repository approach with publishers and service providers.[18]

Initially the project will concentrate on four library and information science journals and PerX, Pilot Engineering Repository Xsearch (see 2.1.2 and 4.2.8).

Meanwhile, OAI-PMH's simplicity may translate into myriad problems for harvesters, especially in cases where data providers do not implement some of the "optional features" that are most helpful to building aggregations. Consequently, service providers are often confronted by inconsistent, insufficient, or incompatible data that limit their ability to build meaningful aggregations for end-users. DLF members and affiliated partners have been at the forefront in articulating these problems and

---

[17] The specification is available from http://www.openarchives.org/OAI/2.0/guidelines-static-repository.htm.
[18] Email communication distribution to the JISC-Repositories listserv on February 21, 2006.

promoting solutions (Cole and Shreeves 2004, Tennant 2004a, Hagedorn 2005b, Shreeves et al. 2005, Lagoze et al. 2006a,b). Their efforts attempt to strike a balance between the demands placed on data providers and the expectations of service providers. To the extent practicable, they seek ways to automate procedures through machine-to-machine interaction, while recognizing that some degree of expert human intervention will always be required.

3.1 Building the Distributed Library

While continuing to apply the lessons learned from adopting the protocol, the DLF received a two-year grant from the Institute of Museum and Library Services (IMLS) in October 2004 to help achieve its vision of "the distributed library," using OAI for digital library aggregation. More concretely, the grant addresses challenges identified by early OAI adopters through promulgating best practices and facilitating communication among data and service providers about such issues as metadata variation, metadata formats, and implementation practices (Shreeves et al. 2005). The proposal states:

> The Open Archives Initiative (OAI) has proven itself as a protocol that allows basic metadata records to be created by many providers and then gathered up by harvesters who use those records to create library services (e.g. http://www.oaister.org/). In the act of using it over several years in library settings, however, a range of issues have come to light that need research and development if OAI is going to mature into its full potential: **collections as well as item records need further development**, and we **need richer mechanisms of creating dialog between harvesters and providers**; the hurdles to adoption need careful study, particularly how to embed the very idea of creating public, harvestable metadata as a **routine step in our digitizing workflows**, and how to **speed up the feedback loop from a harvester to a community of providers** such as exists in the library world, who typically respond positively to such "**good practice**" **guidance**.
> (DLF http://www.diglib.org/architectures/oai/imls2004/, emphasis added)

This is a multi-faceted endeavor, enabling DLF to solidify best practices for the creation of metadata about its dispersed collections, which then inform the development of new DLF services, such as the following:

1. Promotion of the **University of Illinois OAI-PMH Data Provider Registry**, a comprehensive technical resource intended principally for use by builders of OAI services. With nearly 1,050 active repositories, the registry is browsable via a Web interface or as XML. Described more fully in section 4.1.1, the registry is available from http://gita.grainger.uiuc.edu/registry.
2. **A DLF Portal** that allows users to access all items from DLF-member institutions that are publicized through the Open Archives Initiative. Described more fully in

section 4.1.8, as of May 2006, the portal contained more than one million records: http://hti.umich.edu/i/imls/.

3. **A DLF MODS Portal** that represents a subset of the full Portal, gathering together those records that have the richer MODS metadata that support much better subject, date, and geographic navigation. Currently comprising more than 250,000 records, the MODS Portal is also described in section 4.1.8: http://www.hti.umich.edu/m/mods.

4. A new **DLF Collections Registry** that describes nearly 800 publicly accessible digital collections from which the item-level records in the Portal are derived. Described more fully in section 4.4.3, the Registry is available from: http://gita.grainger.uiuc.edu/dlfcollectionsregistry/browse/.

Overall, these efforts are informed by advice from DLF's Scholars' Advisory Panel and Panel of Technical Experts. As a result, Seaman notes that they reflect some of the following improvements:

- A simpler (Google-like) initial interface;
- The inclusion of thumbnail images of graphical collections into the metadata;
- The closing linking of an item to its immediate collection (rather than to its institution);
- More fields for limiting searches;
- A book bag for saving and emailing records;
- Inclusion in A9.com to facilitate simultaneous searching with Amazon, Wikipedia, RedLightGreen, The British Library catalog, and many other OpenSearch services; and
- Persistent URLs.

3.1.1 Components of DLF's Grant-related Work

The DLF grant's principal partners at Emory University, the University of Michigan and the University of Illinois at Urbana-Champaign (UIUC) are focusing on three broad areas of activity:

1. Understanding and improving workflow practices and training so that the creation of item-level metadata is integrated into daily workflow routines of DLF-member institutions. This will increase and regularize the creation and exposure of metadata for harvesting, which serves as the foundation of a reliable and well-populated finding system for DLF's distributed content. Over time, this addresses a chief concern identified in the 2003 survey, namely making the creation and exposure of item-level metadata a priority so more meaningful content is readily accessible to the end-user.[19]

---

[19] http://www.ukoln.ac.uk/repositories/digirep/index/FAQs

2. Developing more nuanced and prescriptive "Best Practices" for the creation of metadata with provisions for richer metadata than the unqualified Dublin Core mandated by OAI. Agreed upon best practices will help to overcome the inordinate amount of time that harvesters (OAI service providers) need to spend normalizing and completing records before they can build new services. This is essential if the proposed finding system is going to scale and flourish. It also addresses concerns of granularity and context raised in the 2003 survey.

3. Coordinating information exchange between service and metadata providers for better discovery by developers as well as by the end-user. This will help digital library developers identify content for new services and will promote wider access by end-users. It responds to the need for a user-friendly registry or discovery tool geared towards the end-user noted in the 2003 survey.

Emory University has created a series of curriculum materials for use in OAI best practices training. This curriculum series includes eight separate documents that together provide a concise set of materials for training institutional teams in best practices for OAI implementation. Emory University is currently developing an online system that would allow searching and collaborative updating of the OAI Best Practices, and controlled output of selected information into formatted training materials. Briefly annotated below, the series is current available at DLF's Web site, http://www.diglib.org/architectures/oai/imls2004/training/:

- **DLF OAI Implementers Workshop: Agenda**
  An example of a typical workshop structure, in which instruction in OAI implementation is paired with a clinical focus on participants' unique circumstances and challenges.
  www.diglib.org/architectures/oai/imls2004/training/OAI-workshopagendaFinal.pdf
- The **Project Abstract**: outlines the purpose and goals of the IMLS project.
  www.diglib.org/architectures/oai/imls2004/training/ProjectAbstractFinal.pdf
- **The Case for OAI**: a synopsis of the background and development of the protocol, motivations for its use (needs and demand), and benefits of increasing the quantity and quality of shared metadata through OAI implementation.
  www.diglib.org/architectures/oai/imls2004/training/CaseforOAIFinal.pdf
- **OAI Implementation: Administrative Planning**: a guide to resource-allocation and planning issues that outlines a six-step implementation scheme and provides the estimated timeline, budget, personnel, and technology.
  www.diglib.org/architectures/oai/imls2004/training/ImplementationFinal.pdf
- **OAI "Cheat Sheet": A Taxonomy of Rapid OAI Deployment Strategies**: provides easy solutions for implementing OAI with current formats and systems and the pros and cons of each strategy.
  www.diglib.org/architectures/oai/imls2004/training/TaxonomyFinal.pdf

- **OAI Tools**: an overview of OAI-implementation tools, grouped into three areas according to their role in the OAI-implementation process: (1) Data Provider Tools & Scripts, (2) Digital Library Systems, and (3) Validation & Harvesting Systems.
  www.diglib.org/architectures/oai/imls2004/training/OAIToolsFinal.pdf
- **Summary of OAI Metadata Best Practices**: introduces (1) the "best practices" for increasing the quality of shareable metadata, (2) the quality issues that currently limit its usability, and (3) the range of metadata formats that may be used with OAI.
  www.diglib.org/architectures/oai/imls2004/training/MetadataFinal.pdf
- **Summary of the DLF Aquifer MODS Profile**: introduces the requirements and recommendations for use with digital cultural heritage materials and humanities-based scholarly resources.
  www.diglib.org/architectures/oai/imls2004/training/MODSFinal.pdf

Most major digital library systems now offer OAI data provider (and increasingly harvesting) capability. The DLF's "OAI Tools" handout annotates seven of them: ContentDM, CWIS (Collection Workflow Integration System), DLXS, DSpace, Ex Libris' DigiTool, Fedora, and Greenstone. While this list concentrates primarily on open source and non-commercial systems, DLF recognizes that many library vendors and software developers are building OAI data provider functionality into their overall digital content management systems. Among major library vendors offering turnkey solutions are: Endeavor's EnCompass; ProQuest's Digital Commons; IndexData's Keystone, Fretwell Downing Informatics' CPORTAL; VTLS' Vortex; and SirsiDynix's 3.0 release of the Hyperion system.

3.1.2 The Case for Sharing Metadata and Improving Its Quality

"Marketing with Metadata – How Metadata Can Increase Exposure and Visibility of Online Content" (Moffat 2006) makes a succinct and persuasive case for exposing metadata by outlining its benefits:

- Allow your content to be found from a large number of locations (e.g. portals, aggregators, search engines).
- Allow aggregators to expose and thereby help to promote your materials in novel ways.
- Enhance the visibility and awareness of your available resources.
- Be a useful way to expose materials to new markets.
- Allow potential users to determine the relevance of resources without having to access them first.
- Facilitate the production of interoperable services.
- Improve the visibility of your content in search engines such as Google, Google Scholar and Yahoo.

- Drive traffic and business to websites.
  (Source: Version 1.0 8th March 2006
  http://www.icbl.hw.ac.uk/perx/advocacy/exposingmetadata.htm)

This advocacy document provides case studies for three ways of exposing metadata: harvesting metadata via OAI, exposing metadata via distributed searching (e.g., Z39.50, SRU/SRW), and exposing content for syndication (e.g., RSS). It also provides answers to common questions, such as:

- **"If it's all about sharing content why can't we just provide you with a link to our content?"**
  Simply providing a link to your content does not allow it to be shared and re-purposed easily and in a standard way. The beauty of exposing metadata in a standard way is that little effort is required for third parties to reuse your metadata and make it available to their visitors.
- **"I don't like the thought of giving away our content for others to use."**
  Exposed metadata usually only contains a brief description of the actual content - just enough to generate interest in potential users. These users will be directed back to your site by links in the metadata in order to access the full content in the normal way (i.e. freely available, subscription based, pay-per-view, etc).
- **"Will exposing my metadata mean that it is indexed by search engines such as Google or Google Scholar?"**
  This depends on how your metadata is exposed and the indexing approaches taken by individual search engines. Exposing metadata via OAI certainly can improve ranking in search engines. "A normal Google or Google Scholar search favours OAI-repository material and normally ranks it higher than an individual's own website." Recent developments such as 'search engine-OAI bridges' are improving search engines indexing of OAI compliant repositories. Many OAI repositories are now indexed by a number of search engines, e.g. Cogprints, a repository for cognitive sciences, is indexed by Google, Google Scholar, Yahoo, Scirus and Citebase.
- **"Why can't I simply make my content available to Google and let people find my stuff that way?"**
  You can, and in many cases this will be a perfectly appropriate thing to do. This is particularly true for freely available full text resources. However, in some cases, for example where most of your resources are not text-based, exposing them to Google may not help much. In other cases, you may not want to make the full content freely available. In these situations, exposing metadata may be more appropriate. By making your metadata freely available, you can allow people to discover your resources more readily.

The case made, it is no surprise given their extensive experience in building services based on OAI harvesting, that the DLF in partnership with the National Science Digital

Library (NSDL) is developing "Best Practices for Shareable Metadata." They point out that the attributes of high-quality metadata in a local context do not necessarily equate with the best metadata in a shared environment. The guidelines identify the following additional desirable characteristics for shareable metadata:

- **Proper context**. *In a shared environment, metadata records will become separated from any high-level context applying to all records in a group, and from other records presented together in a local environment. It is therefore essential that each record contain the context necessary for understanding the resource the record describes, without relying on outside information.*
- **Content coherence**. *Metadata records for a shared environment need to contain enough information such that the record makes sense standing on its own, yet exclude information that only makes sense in a local environment. This can be described as sharing a 'view' of the native metadata.*
- **Use of standard vocabularies**. *The use of standard vocabularies enables the better integration of metadata records from one source with records from other sources.*
- **Consistency**. *Even high-quality metadata will vary somewhat among metadata creators. All decisions made about application of elements, syntax of metadata values, and usage of controlled vocabularies, should be consistent within an identifiable set of metadata records so those using this metadata can apply any necessary transformation steps without having to process inconsistencies within such a set.*
- **Technical conformance**. *Metadata should conform to the specified XML schemas and should be properly encoded.* (DLF and NSDL 2005)

Since August 2005, a working draft of the "Best Practices" has been available at the project's wiki located at the NSDL's community portal. DLF and NSDL are working with their respective communities and library vendors alike to raise awareness of why these guidelines are important. Publication of the document is anticipated later in 2006.

A key recommendation emanating from this collaboration is to endorse MODS (Metadata Object Description Schema) as the preferred metadata schema, particularly for use in describing cultural heritage and humanities digital resources.[20] In December 2005, the DLF released for public comment, "MODS Implementation Guidelines for Cultural Heritage Materials." Among other features, the guidelines help to address the particular difficulties inherent in describing digital objects that have analog originals by distinguishing between "the intellectual content and genre of a resource and its digital format and location." The richer MODS descriptive schema helps to pave the way for enhanced service features, such as those identified by DLF's Scholars' Advisory Panel.

---

[20] The Library of Congress Web site about MODS is available from
http://www.loc.gov/standards/mods/.

MODS high-level elements include:

• Title Information	• Note
• Name	• Subject
• Type of resource	• Classification
• Genre	• Related item
• Origin Information	• Identifier
• Language	• Location
• Physical description	• Access conditions
• Abstract	• Extension
• Table of contents	• Record Information
• Target audience	(Guenther 2005)

Institutions currently creating MODS records include: the Library of Congress, Indiana University, OCLC, and the University of Chicago (refer to section 4.1.8 for sample records).

3.1.3 DLF 2006 Survey Responses about Metadata

The DLF survey respondents represent primarily OAI service providers so it comes as no surprise that most are eagerly anticipating promulgation of the "best practices" to improve the quality of metadata they harvest. Indeed representatives from a number of these services are members of the DLF Best Practices Task Force. When asked if they expected to change their metadata creation practices in light of the forthcoming DLF/NSDL OAI best practice guidelines, the two respondents below reflect the hopes of most service providers:

> *We expect the uncertainly of metadata normalization and enhancement that we have to do to lessen as better standards/guidelines are promulgated for mapping native content management metadata into OAI records for harvesting.*

> *We would like to incorporate alternate metadata formats, besides oai_dc, whenever possible. We would also like to incorporate collection descriptions into the search interface when feasible. However, as an aggregator we are pretty much stuck with whatever metadata is available from our sources. We do some date normalization and will follow best practices as applicable; however, our hope is that the best practices will influence the repositories from which we harvest so that we can take advantage of the improved metadata to provide better search and browse services.*

Although they are at the forefront of devising tools that help to migrate, remediate, and enhance metadata, these service providers also join other survey respondents calling for more automatic metadata tools.

Respondents noted some of the following accomplishments relative to metadata:

- Emory University created The Metadata Migrator software package, funded by the Institute of Museum and Library Services. It allows institutions such as museums, archives, research centers, and small libraries to make their locally stored records available for online searching using the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH), http://www.metascholar.org/sw/mm/.
- The California Digital Library drafted "Specifications for Metadata Processing Tools" (Tennant, n.d.), http://www.cdlib.org/inside/projects/harvesting/metadata_tools.htm, and created a Date Normalization Tool. This Java utility takes non-machine readable Common Era dates as input and outputs machine-readable dates in order to enhance digital collections to support date range queries. Through its Metasearch Initiative, CDL established an SRU-compliant gateway to OAI-harvested metadata. These initiatives are described at http://www.cdlib.org/inside/projects/harvesting/.
- The CIC launched a new consortial metadata portal under the leadership of the University of Illinois; drafted "CIC-OAI Project Recommendations for Dublin Core Metadata Providers," Version 1.0 (06/18/2004), edited by Muriel Foulonneau and Timothy W. Cole, http://cicharvest.grainger.uiuc.edu/dcguidelines.asp; and refined its workflow and filtering processes for metadata as described at http://cicharvest.grainger.uiuc.edu/aggregation.asp. It enriched and normalized the metadata to support various browse and search interfaces. Recently it developed a new enhanced OAI data provider for the registry to allow not only simple Dublin Core records that describe each repository harvested, but also the much richer collections that created manually along with the repository descriptions imported from OAIster. Because each of these sets and subsets has rich collection-level metadata derived from the registry, it allows harvesters to associate collection-level metadata to individually harvested items more easily.
- OAIster provided UI with additional metadata about all of its OAI repositories (e.g., title, description, home page, and historical record counts) and now it refers new data providers to UI for registration and validation before harvesting their metadata. In March 2006, OAIster announced the availability of its metadata for use by federated search engines via SRU and created a Web page with instructions about how it use its metadata outside OAIster's interface (http://oaister.umdl.umich.edu/o/oaister/sru.html).
- The DLF launched a new portal based on MODS (as 4.1.8), http://www.hti.umich.edu/m/mods.
- The Directory of Open Access Journals (DOAJ) began to make article-level metadata available in addition to journal metadata, http://www.doaj.org/articles/questions#metadataA.
- The Open Language Archives Community (OLAC) developed a metadata quality, report system and implemented an interactive survey of OLAC

metadata implementations that permits users to see how any attribute or field of OLAC metadata is used by other OLAC archives. Its new search engine includes in the results a metadata quality-centric sorting algorithm.

- SMETE resources are cataloged to meet the requirements of the IEEE Learning Object Metadata Standard and SMETE has developed tools to transform local application profiles to normalized application profiles, http://smete.org/smete/ (see Technology).
- BEN has created metadata validation software tools for contributors to its portal, http://www.biosciednet.org/project_site/.
- DLESE developed a distributed Web-based cataloging tool to support multiple collections and multiple metadata frameworks. With other entities created the ADN Metadata Framework, http://www.dlese.org/Metadata/adn-item/history.htm.
- MERLOT developed a Metadata Services Agreement for use with participating external vendors.
- Cornucopia migrated to a new software system and realigned almost all of its data structure to conform to the RSLP (Research Support Libraries Programme) Collection Level Description Metadata Schema, http://www.ukoln.ac.uk/metadata/rslp/.
- In creating the collection registry model, the IMLS Digital Collections & Content (DCC) gateway draws on research about how to define and describe collections, ultimately opting to adapt the RSLP Collection Description Schema and the Dublin Core Collection Description Application Profile (Cole and Shreeves 2004, 312). The DCC arrived at a collection description metadata schema with four classes of entities
- The Collaborative Digitization Program (e.g., Heritage West) revised and updated its CDP Dublin Core Metadata Best Practices, http://www.cdpheritage.org/cdp/documents/CDPDCMBP.pdf.
- The American West carried out preliminary work on metadata enhancement to support topical clustering and faceted browsing.
- DLF Aquifer developed a descriptive metadata (MODS) profile (as described above), http://www.diglib.org/aquifer/DLF_MODS_ImpGuidelines_ver4.pdf. Next steps will be developing middleware tools that support metadata management activities such as migration, taxonomy assignment, and metadata enrichment.
- The INFOMINE database has been populated with robot records, mostly created from the iVia virtual library crawler and machine-generated metadata (using iVia classifiers). The iVia software supports automated metadata generation to assign Library of Congress Subject Headings and LC Classifications to resources. The iVia software also enabled NSDL to harvest item-level metadata from iVia's server for selected NSDL collections that did not include detailed metadata.

Among the challenges, respondents noted:

- the willingness (or not) to make harvestable metadata a local priority (DLF Portal);
- lack of a good metadata editor and metadata cleansing tools (OLAC);
- quality control of metadata and learning objects (NEEDS);
- incompatible metadata standards (Sheet Music Consortium);
- automatic metadata creation tools (Intute); and
- the need for robust, flexible, open source tools for metadata normalization and enrichment (CDL).

Finally, turning to the goals of "next generation" services, the Sheet Music Consortium seeks enriched metadata to provide better retrieval services and INFOMINE looks forward to harvesting and sharing metadata with other digital libraries. Meanwhile, NSDL's conversion to a Fedora repository marks a major transition from a metadata-centric to a resource-centric data model and search service; and DLF Aquifer anticipates experimentation with methods of aggregation other than metadata harvesting, namely the ability to move digital objects from domain to domain, perhaps modifying and re-depositing them in a different location in the process.

3.2   Digital Library Services Registries (DLSR)

The "service registry" is central to enabling digital libraries to interoperate in distributed information environments with service-oriented architectures based on common standards and protocols. Dempsey refers to the service registry's collective data as the equivalent of the "systemwide 'intelligence'" within a network to run distributed applications.[21] At a March 2006 workshop, participants wrote a draft definition of the concept: "A digital library service registry allows a machine or human to discover available digital library services, locate those services, and obtain configuration information to services for the purpose of interfacing."[22]

While there are numerous examples of application-specific registries like those for OAI-PMH (described in the next section) or the OpenURL Registry, maintained by OCLC, two efforts underway in the UK and US respectively share the goal of creating "service

---

[21] Discussed in Lorcan Dempsey's Web blog, "From Metasearch to Distributed Information Environments," reporting on presentations given at the NISO OpenURL and Metasearch meeting. Posted on October 9, 2005; available from http://orweblog.oclc.org/archives/000827.html.

[22] Refer to the workshop Web site and the report prepared thereafter, available from http://wiki.library.oregonstate.edu/confluence/display/DLSRW/Home. Presentations are located at http://wiki.library.oregonstate.edu/confluence/display/DLSRW/WorkshopPresentations.

agnostic" registries to facilitate resource discovery and use by a multitude of applications (e.g., Z39.50, SRW/U, OAI-PMH, OpenURL).[23] The Information Environment Service Registry (IESR, http://www.iesr.ac.uk/) sponsored by JISC (described in section 2.1.1 above) uses a centralized approach (depicted as part of the "shared infrastructure" in Figure 06), whereas the Ockham DLSR (http://www.ockham.org/) sponsored by the NSF's National Digital Science Library (NSDL) relies on a distributed model.

The DLSR serves three primary functions:

- *Discovery* - allowing a user or a machine to discover available, relevant services;
- *Resolution* – providing the ability for a person or a machine to locate, or resolve to, a
- *Configuration* – provide information necessary for a client to access a particular service. (Frumkin 2006a, 24)

As part of the DLSR Workshop noted above, Frumkin devised a series of "use cases" that demonstrate how DLSR fits into different scenarios. In the two examples below, the first depicts how the DLSR would help a user through personalized metasearching and the second, illustrates how the DLSR would facilitate development of OAI aggregations.

Personalized Metasearching:
Bernie the researcher is exploring the history of science for a book he is working on. He uses his library's metasearch tool to search through history-related databases and collections. He also searches the web for collections and resources that do not reside within the context of his library's collections and services. During his web searching, he discovers the Linus Pauling papers at Oregon State University. Bernie would like to add the Linus Pauling collection to his default metasearch, so he goes into the metasearch tool, clicks the 'customize this search' button, and then searches for the Linus Pauling collection to see if he might be able to add it. The metasearch application searches the digital library service registry for the Linus Pauling collection, and shows Bernie that there are actually five collections which match his search. Bernie is delighted to find four more collections, and checks all of them to be added. The metasearch application then discovers that these collections are not immediately searchable via any standard protocol, but they are harvestable via the OAI-PMH protocol. The metasearch tool is intelligent enough to be able to automatically start harvesting the metadata from these collections into a local index, and then include them as part of Bernie's default search.

Name Authority Identification
Jim is in charge of setting up an OAI-PMH aggregator that will gather distributed metadata records and then reuse them in a science digital library. He is concerned about

---

[23] Available from http://www.oclc.org/americalatina/en/research/projects/openurl/registry.htm.

the quality of the collected records and would like to apply some normalization and cleanup to them. One particular area of concern is the uncontrolled use of personal and corporate names in the records. He uses the service registry to locate existing name authority services offered by various organizations, and plans an aggregation strategy that uses these services for metadata cleanup.
Source: wiki.library.oregonstate.edu/confluence/display/DLSRW/RegistryUseCases

4.0   Review of Resources

4.1   Points of Reference: Open Access and the Open Archives Initiative

While OAI and Open Access are not synonymous, the Open Access movement relies heavily on the OAI protocol as the mechanism for communicating the availability of OA resources. Publishing in Open Access journals and self-archiving in OA archives are specified by the Budapest Open Access Initiative (BOAI), and further bolstered by the Berlin Declaration, as the major ways to make manifest OA research output. Moreover, institutional repositories (typically OA and OAI-compliant) are increasingly accepted as an essential component of a university's scholarly infrastructure (Lynch 2003).

When the 2003 report was written, it was difficult to identify OA (and OAI-compliant) journals and repositories. The Directory of Open Access Journals (DOAJ), launched in May 2003, marked an initial step towards making OA journals better known, but it was still in an early stage of development. In addition, there was no easy way for authors to identify the copyright polices and self-archiving regulations of publishers. Discovery of OA repositories was even more problematic. Assembling a composite picture was painstaking and idiosyncratic, made possible only by triangulating from data gathered from multiple sources—the official Open Archives Initiative's voluntary registry of OAI data and service providers, the technical OAI Repository Explorer validation system, and via the aggregators, such as Arc and OAIster.

Noting numerous difficulties in identifying OAI-compliant repositories and the deleterious impact on data providers, service providers and their users, the 2003 DLF report called for a user-friendly comprehensive registry (Brogan 2003, 75). In the intervening years, the situation has changed markedly. The registries, directories and indexes under consideration here, are the visible manifestation of OA and OAI growth.

New to the scene, the University of Illinois OAI-PMH Data Provider Registry now ably serves as a comprehensive interactive OAI identification system. Indeed, from a technical standpoint, the concept of registries has become an essential component of digital library architecture covering a wide spectrum of functions. Two new projects,

one in the UK and the other in the US, are working in tandem to develop a framework for DL service registries that will help to automate the discovery of DL content and services (see section 2.1.1 and 3.2). Meanwhile, two OA repository registries, geared towards improving communication between developers, researchers and authors, have been developed in the UK. Concomitantly, the DOAJ has extended its services to include article-level access and two new directories now monitor journal and publisher copyright and self-archiving permissions (one outcome of the UK's RoMEO project cited in 2003). Arc continues to serve as a test bed for improving and extending OAI applications. OAIster, on the other hand, has become the de facto leader as a global OAI service provider, dispensing item-level digital content to end-users.

In addition to discussing these major services, this section reviews three new consortial metadata aggregators—the CIC Metadata Portal and the DLF's OAI and MODS portals and then turns to Germany as an exemplar of nationally-based OAI services. Critical issues and future directions round out the review of these services.

Interlocking Purposes

Collectively, the registries, directories and indexes under review serve the following purposes for an audience ranging from data and service providers to researchers, authors and end-users:

- Raise awareness and visibility within the technical community so digital resources (or metadata) are publicized and harvested.
- Offer technical validation systems to test OAI-PMH conformance.
- Serve as a test bed for research and development to improve future OAI services.
- Improve communication between data and service providers.
- Provide mechanisms for the developer community to stay current (through email forums or RSS feeds).
- Promote Open Access principles and promulgate institutional policies adhering to the BOAI and Berlin Declaration.
- Publicize repositories upholding OA principles.
- Monitor the status, growth and function of OA implementations across time, country, type of media and software.
- Inform authors of OA journals or repositories where they can publish (or self-archive) their research output, thereby increasing its impact.
- Inform authors of institutional or journal policies pertinent to self-archiving or copyright permissions.
- Serve as a comprehensive directory of OA institutional participants and a feedback loop for constituents from developers to end-users.
- Provide end-users with full-text article or digital object-level access to academic resources in a timely way through reliable services.
- Monitor the impact of OA and OAI adoption and use.

Table 05: Summary of General OA and OAI Services: Size, Goal, and Core Audience (March 14, 2006)

TECHNICAL REGISTRIES	
Open Archives Initiative http://www.openarchives.org/ 404 data providers and 23 service providers	Official, voluntary registry of OAI data and service providers to facilitate awareness, technical compliance, and community participation. Core Audience: Developers
University of Illinois OAI-PMH Data Provider Registry http://gita.grainger.uiuc.edu/registry/ 1,047 repositories (955 actively responding)	Comprehensive interactive registry and database of OAI implementations for discovery, technical perusal, and community development. Core Audience: Developers
DIRECTORIES OF OA JOURNALS AND SELF-ARCHIVING POLICIES	
DOAJ: Directory of Open Access Journals (DOAJ) http://www.doaj.org/ 2,113 journals of which 567 provide access to 90,710 articles	Authoritative, comprehensive directory of scholarly journals adhering to BOAI open access principles with growing body of article-level access. Core Audience: Service providers (libraries, aggregators, metadata harvesters), researchers, and authors.
Publisher Copyright Policies & Self-Archiving: SHERPA/RoMEO List http://www.sherpa.ac.uk/romeo.php 135 publishers and circa 9,000 journals	List of copyright and self-archiving policies of scholarly journals and publishers. As part of SHERPA, the British Library provides current information about the link between publishers and particular journal titles. Core Audience: Authors
Journal Policies—Self-Archiving Policies of Journals http://romeo.eprints.org/ 129 publishers and 8,698 journal titles	Directory of scholarly journals and publisher self-archiving policies extracted from SHERPA/RoMEO data. Extensive distinctive statistical data with literature citations in support of self-archiving and OA publishing to strengthen impact of research output. Audience: Authors

DIRECTORIES OF OA REPOSITORIES	
ROAR: Registry of Open Access Repositories http://archives.eprints.org 640 archives from 40 countries and 3,728,201OAI records (from 480 archives in Celestial)	Registry to monitor overall growth in the number of e-print archives and maintain a list of GNU EPrints sites. Core Audience: E-print community of developers and researchers.
ROARMAP: Registry of Open Access Repository Material Archiving Policies http://www.eprints.org/openaccess/policysignup/ 18 policies from 9 countries plus 1 European research agency	Directory of institutions with self-archiving policies with associated deposit growth charts, model statements and rationale in support of BOAI and Berlin Declaration. Core Audience: OA research community.
OpenDOAR: Directory of Open Access Repositories	Comprehensive and authoritative list

(under development)	of institutional, subject- and funder-
http://www.opendoar.org/	based repositories.
355 repositories	Core Audience: Developers and
	researchers.
CROSS-ARCHIVE SEARCH SERVICES AND INDEXES	
Arc: Cross-Archive Search Service	The first implementation of a
http://arc.cs.odu.edu/	hierarchical OAI harvester
177 archives and more than 7 million records	(aggregator) serves as a cross-archive
	search service and R&D test bed to
	improve OAI services.
	Core Audience: Developers
OAIster	Search and discovery service providing
http://www.oaister.org/	access to OAI item-level digital objects,
597 institutions from 42 countries and more than 7	including some licensed restricted
million records	access materials.
	Core Audience: Researchers

Technical registries serving a range of purposes are rapidly becoming key components of the standards and technology infrastructure supporting digital libraries. Facilitating interoperability through a low-barrier protocol, the official Open Archives Initiative site does not require either data or service providers to register in order to implement the protocol. Registration is optional and many developers simply do not take the time. In other instances they may deliberately choose not to register because the service is not yet in full production; they do not wish to publicize the availability of their resources; or they already have a known clientele. Registration, however, is not merely a matter of publicizing a new repository or service; it also typically entails testing archives for compliance with the OAI protocol. This helps to validate that metadata is appropriately configured to meet at least minimal standards for harvesting. OAIster, for example, requests new data providers to follow a series of steps before contacting them to harvest new content. OAIster's guidelines include official registration with the Open Archives Initiative where new data providers can obtain the OAI foundational documents, access basic OAI tools, and join community services, consisting of email forums and other registries for data and service providers. As a final step prior to contacting OAIster, new data providers are asked to email the administrator of the University of Illinois's Registry (described below) thereby helping to ensure that it has a complete listing of OAI repositories. OAIster's implementation steps help to reinforce the role and function of different OAI registries and validating services, leading to a more cohesive community of practice.

4.1.1 University of Illinois OAI-PMH Data Provider Registry

Developed under the auspices of the DLF's IMLS National Leadership Grant described earlier in this report, this registry primarily serves as a tool for OAI harvesters to discover and effectively use content in repositories upon which developers can build services. The UI Registry (announced in October 2003) strives to be comprehensive and deploys a systematic multi-faceted approach (that goes beyond self-registration) to achieve the goal of completeness (Habing et al. 2004; Shreeves et al. 2005). As of mid-May 2006, the Registry, with 1,042 repositories, is the most complete and useful OAI data provider discovery service for developers.

It automatically harvests an array of data elements from each repository, making "it possible to search for OAI repositories using various criteria and browse through different views of the registry [e.g., sets, metadata formats, records, identifiers, subjects] without any manual cataloging of the various OAI repositories" (Shreeves et al. 2005, 581). A new enhanced OAI data provider has been developed for the registry to allow not only simple Dublin Core records which describe each repository to be harvested, but also the much richer information that has been created manually along with the repository descriptions imported from OAIster (Cole and Habing 2006).[24] The metadata format for these richer descriptions conforms to the schema developed for UIUC's IMLS Digital Collections & Content project (see section 4.4.2). UIUC has also developed an OAI gateway application that provides a single point of harvest for all DLF-member repositories. Beyond the convenience of harvesting from a single base URL, individual repositories are organized as sets within the gateway with their own sets organized as subsets. Because each of these sets and subsets has rich collection-level metadata derived from the registry, it allows harvesters to easily associate collection-level metadata to individually harvested items. The DLF member OAI data providers are cataloged and browsable by GEM (Gateway to Education Materials) and LCSH (Library of Congress Subject Headings).

The UI Registry and OAIster collaborate to improve communication between OAI data and service providers, while also enhancing their respective services. Initially OAIster provided UI with additional metadata about all of its OAI repositories (e.g., title, description, home page, and historical record counts) and now it refers new data providers to UI for registration and validation before harvesting their metadata. This helps to ensure fuller coverage via the UI Registry while also resolving some technical validation problems prior to harvesting by OAIster. OAIster also sends its historical data to the Registry on a monthly basis. This makes it possible to access growth graphs for many repositories, although it does not match ROAR's growth charts in terms of user-friendliness and access. The Registry's syndication service (RSS) alerts users to recent

---

[24] Information about recent work underway is taken from the DLF's Interim Report (October 2005-April 2006) submitted to IMLS in support of its National Leadership Grant.

changes, listing modifications and new additions over the past 30 days. In addition to OAI-PMH and RSS export functionality, it also supports the SRU protocol (CQL subset). UI is also developing Web-based search and browse interfaces for an OAI service provider registry that will list services developed from harvesting data via the OAI-PMH. Eventually, UI hopes to link the OAI service providers in the database to the OAI data providers from which they harvest. Project news, presentations, and documents, including the cataloging procedures and guidelines used for the DLF collections is available at the Registry's Web site.

4.1.2 DOAJ: Directory of Open Access Journals

Update Table 01: DOAJ based on DLF Survey responses, Fall 2005

	DOAJ (Directory of Open Access Journals) http://www.doaj.org/
ORGANIZATIONAL MODEL	Hosted, maintained and partly funded by Lund University Libraries Head Office. Other current sponsors: Open Society Institute, SPARC Europe, BIBSAM (National Library of Sweden), Axiell AB.
SUBJECT	Cross-disciplinary
FUNCTION	Covers free, full-text, quality-controlled scientific and scholarly journals. All subjects and languages.
PRIMARY AUDIENCE	Service Providers, Research Community
STATUS	Established
SIZE	1,909 journals of which 467 are searchable at article level, comprising 80,687 articles.
USE	No response
ACCOMPLISHMENTS	1. Article metadata search. 2. Journal owner admin functions. 3. OAI-harvesting on both journal and article level.
CHALLENGES	1. Add more content. 2. Include OA articles from hybrid journals
TOOLS OR RESOURCES NEEDED	No tools needed.
GOALS OF NEXT GENERATION RESOURCE	Dissemination

Launched in May 2003 with 350 journals, DOAJ included more than 1,900 titles by December 2005 and quickly surpassed the 2,000 mark in early 2006. Article-level searching was introduced in June 2004 and as of mid-March 2006 exceeded 80,000 articles. According to the DOAJ Web site, "The Directory aims to be comprehensive and cover all open access scientific and scholarly journals that use a quality control system to guarantee the content." It defines open access journals as those that "use a funding model that does not charge readers or their institutions for access" and its selection criteria uphold reader's rights as put forward in the BOAI principles to "read, download,

copy, distribute, print, search, or link to the full texts of these articles." In early 2006, DOAJ updated its selection criteria based on feedback from users.

DOAJ Selection Criteria

Coverage:
- Subject: all scientific and scholarly subjects are covered.
- Types of resource: scientific and scholarly periodicals that publish research or review papers in full text.
- Acceptable sources: academic, government, commercial, non-profit private sources are all acceptable.
- Level: the target group for included journals should be primarily researchers.
- Content: a substantive part of the journal should consist of research papers. All content should be available in full text.
- All languages.

Access:
- All content freely available.
- Registration: Free user registration online is acceptable.
- Open Access without delay (e.g. no embargo period).

Quality:
- Quality control: for a journal to be included it should exercise quality control on submitted papers through an editor, editorial board and/or peer-review.

Periodical:
- The journal should have an ISSN (International Standard Serial Number, for information see http://www.issn.org/).
  (Source: http://www.doaj.org/articles/questions#definition)

Table 06: DOAJ Journal Subject Coverage (March 6, 2006)

	DOAJ Titles N = 2,081	Percent of Total
Agriculture and Food Sciences	122	5.9%
Arts and Architecture	52	2.5%
Biology and Life Sciences	231	11.1%
Business and Economics	56	2.7%
Chemistry	52	2.5%
Earth and Environmental Sciences	159	7.6%
General Works-Multidisciplinary	26	1.2%
Health Sciences	711	34.2%
History and Archaeology	93	4.5%
Languages and Literatures	120	5.8%
Law and Political Science	90	4.3%
Mathematics and Statistics	95	4.6%
Philosophy and Religion	71	3.4%
Physics and Astronomy	77	3.7%

Science General	6	0.3%
Social Sciences	483	23.2%
Technology and Engineering	159	7.6%

The DOAJ subject classification is expandable and offers links from topical categories to the journal titles. The two largest sub-categories are Medicine (General) with 194 titles (in Health Sciences) and Education with 148 titles (in Social Sciences). Users can search for journals via keywords or browse by title or subject. The article database supports basic Boolean operators to connect keyword or phrase searches across all fields or limited to title, journal title, author, ISSN, keyword or abstract. A search for articles using the keyword <tsunami> retrieves 13 documents, all with 2005 and 2006 publication dates. The entries provide basic bibliographic citations with the option to view the record or the full text article.

Information about harvesting DOAJ journal and article-level metadata (initiated in July 2004) as well as restrictions on metadata usage (DOAJ is licensed under the Creative Commons Attribution-ShareAlike License) is provided at the Web site's FAQ. DOAJ supports harvesting of broad subject-based sets. DOAJ actively solicits monetary contributions from users to continue to improve its functionality and keep it in continuous operation.

4.1.3 Directories of Journal and Publisher Copyright and Self-archiving Policies

While DOAJ identifies Open Access journals and publishers it does not disclose their copyright or self-archiving policies. Authors can use the SHERPA/RoMEO List of Publisher Copyright Policies and Self-archiving to "find a summary of permissions that are normally given as part of each publisher's copyright transfer agreement."[25] The directory, hosted by the University of Nottingham, is searchable by journal title or publisher. Publishers are assigned a color code that reflects whether permission is granted to self-archive and at what stage in the publication process. According to the site's summary statistics in May 2006, 78 percent of the 154 publishers officially allow some form of self-archiving. An API is being developed to allow repository administrators and others to interface with the database, possibly as a stage in a repository's ingest procedure or similar process. The information is available for downloading by interested

---

[25] SHERPA stands for Securing a Hybrid Environment for Research Preservation and Access. The initiative embraces various projects including OpenDOAR (described below), led by the University of Nottingham with funding from JISC and CURL. Additional information is available from http://www.sherpa.ac.uk/index.html It builds on the work of Project RoMEO (Rights Metadata for Open Archiving): http://www.lboro.ac.uk/departments/ls/disresearch/romeo/.

parties by special arrangement: for example, the listing hosted by Eprints.org is based on the SHERPA/RoMEO information. Reports and publications emanating from SHERPA affiliated projects and research are available from its Web site, http://www.sherpa.ac.uk/guidance/advocacy.html#reports.

Table 07: Statistics for the 135 publishers on SHERPA/RoMEO (March 2006)

RoMEO color	Archiving policy	Publishers	%
Green	Can archive pre-print and post-print	59	44
Blue	Can archive post-print (i.e. final draft post-refereeing)	30	22
Yellow	Can archive pre-print (i.e. pre-refereeing)	14	10
White	archiving not formally supported	32	24

Source: http://www.sherpa.ac.uk/romeo.php?stats=yes (March 18, 2006)

EPrints.org has developed a similar directory, based on SHERPA/RoMEO's data of journals that have and have not already given "their green light to author self-archiving." Under rapid development, as of mid-March 2006, it contains 136 publishers and almost 8,900 journals. In contrast to the SHERPA/RoMEO's list, journals are given one of three different color codes:

- Green: Permits post-print self-archiving
- Pale Green: permits preprint self-archiving
- Grey: Does not permit self-archiving

The site maintains summary statistics by journal as well as publisher. Amalgamating green and pale-green publishers results in 76 percent of publishers officially permitting self-archiving (the equivalent of SHERPA/RoMEO's green plus blue plus yellow publishers). In contrast to SHERPA's list, however, the EPrints.org site also provides the data based on journal titles, resulting in a much higher percentage of self-archiving permission rate: 93 percent of the 8,265 journals listed "green" (69 percent full green and 24 percent pale green). A more detailed statistics page highlights and updates the findings of seminal studies about self-archiving (Swan and Brown, 2004a,b; 2005; Harnad et al. 2004; and Harnad and Brody 2004) with charts depicting the current proportion of toll-access and OA articles and the current potential for immediate OA provision.[26]

Comparative Coverage: OA Journal Directories and Databases

---

[26] Available from http://www.ecs.soton.ac.uk/~harnad/Temp/Romeo/romeosum.html .

It seems reasonable to expect that SPARC's OA journal titles would be well-represented in these OA journal directories, but a comparison of sample SPARC journal titles reveals inconsistent and incomplete coverage.

Table 08: SPARC Open Access Journals Represented in DOAJ, PubMed Central, SHERPA/RoMEO, EPrints.org List and EZB (March 18, 2006)

SPARC OPEN ACCESS JOURNALS	DOAJ	PubMed Central	SHERPA/ RoMEO Publisher Policies	EPrints.org Journal Policies	EZB
Documenta Mathematica	Journal only	N/A	Not listed	Not listed	Green
Economics Bulletin	Journal only	N/A	Not listed	Not listed	Green
Geometry & Topology and Algebraic & Geometric Topology	Not listed.	N/A	Title only	Not listed	Green
Journal of Insect Science	Yes, with content.	Immediate free and OA without delay.	Title only	Not listed	Green
Journal of Machine Learning Research	Journal only	N/A	Yellow	Pale-Green	Green
New Journal of Physics	Journal only	N/A	Not listed.	Not listed	Green
Optics Express	Journal only	N/A	Not listed.	Green	Green
PLoS Biology PLoS Computational Biology PLoS Genetics PLoS Medicine PLoS Pathogens	Yes, all titles, with content.	Immediate free and open access without delay.	Green Not listed Not listed Title only Not listed	Green Not listed Not listed Green Not listed	All 5 Green
BioMed Central	Yes, with majority of content.	Immediate free and OA without delay except for five titles with 24-month delayed access to non-research articles.	Green	Green: 144 journals	Green

Project Euclid Journals 6 OA titles (25 titles partially open after 3-5 years and 9 titles by subscription)	Of 6 OA titles: 4 not listed; 2 journal only.	N/A	Of 6 OA titles: 1 Green; 1 title only; 4 not listed.	Of 6 OA titles: 2 Green, 4 not listed.	All 6 Green

All of the PubMed Central titles indicated as free and open access by SPARC also have article-level access in DOAJ. However, with the exception of the fully-represented BioMed Central titles, coverage of other PMC titles is uneven in the two self-archiving policy directories. Of the six OA Project Euclid journal titles, only two are listed in DOAJ; one title is identified as green in SHERPA and two in EPrints.org. Among the sample OA titles: 5 are not listed in DOAJ; fifteen are either not listed or represented by title only (without any corresponding self-archiving policy information) in SHERPA; and twelve are not covered in EPrints.org. The German database of e-journals, EZB (Elektronische Zeitschriftenbibliothek), is the only source to contain all of SPARC's OA titles; moreover, they are correctly annotated in cases where only specific years are OA. EZB's coverage and coding scheme is described more fully below (see 4.1.9) but it does not include journal or publisher self-archiving policies.

4.1.4 ROAR: Registry of Open Access Repositories

Launched in fall 2003, ROAR (formerly known as the Institutional Archives Registry) has two main functions: "(1) to monitor overall growth in the number of e-print archives and (2) to maintain a list of GNU EPrints sites (the software the University of Southampton has designed to facilitate self-archiving)."[27] The ROAR FAQ lays out the goals for coverage, emphasizing OA and OAI-compliant research documents, predominantly preprints, postprints of peer-reviewed journal articles, or dissertations. In practice, it has few, editorial exclusions.[28]

Beyond research papers, ROAR includes other formats; for example, the University of Southampton's Crystal Report Structure Archive (http://ebank.eprints.org/), a repository that utilizes EPrints.org software to archive datasets "generated during the course of a structure determination from a single crystal x-ray diffraction experiment." It also includes records (46,000) from the Biblioteca "Dr. Jorge Villalobos Padilla, S.J." Instituto Tecnológico y de Estudios Superiores de Occidente, (ITESO), Mexico, excluded by

---

[27] As cited in the FAQ available from http://trac.eprints.org/projects/iar/wiki/FAQ .

[28] In an announcement about ROAR posted by Stevan Harnad to the AmSci Forum, he states that "archives that do not provide *any* full-text content at all (only metadata), or that provided only content of other kinds (internal documents, courseware, library records, audio video, software) are not covered—though archives of *mixed* content (both OA and non-OA) are covered." As of mid-May, this level of editorial control has not been applied to ROAR.

OAIster because they report that many items refer to SFX links, hence they are not really OA. As stated elsewhere in this report, there are many "grey" areas in OAI-harvesting that make it difficult to reach uniform decisions about such parameters as "freely available" or "Open Access."

ROAR is a useful tool for analyzing the characteristics, size, and growth within and across OA e-print archives around the world. Archives are classified by country, system software, and content type. Searches can be filtered by any combination of these fields (e.g., Research Cross-Institution archives using DSpace in Belgium) and sorted by Name, Datestamp, or Total OAI Records. Results provide an annotated entry about the resource with links to the source site, an estimate of the percent of its content that is freely accessible, full text summary graphs charting its growth over time, and a thumbnail of the service's Web site.

Figure 08: Screenshot of ROAR sample search result for Belgium, DSpace, Research Cross-Institution
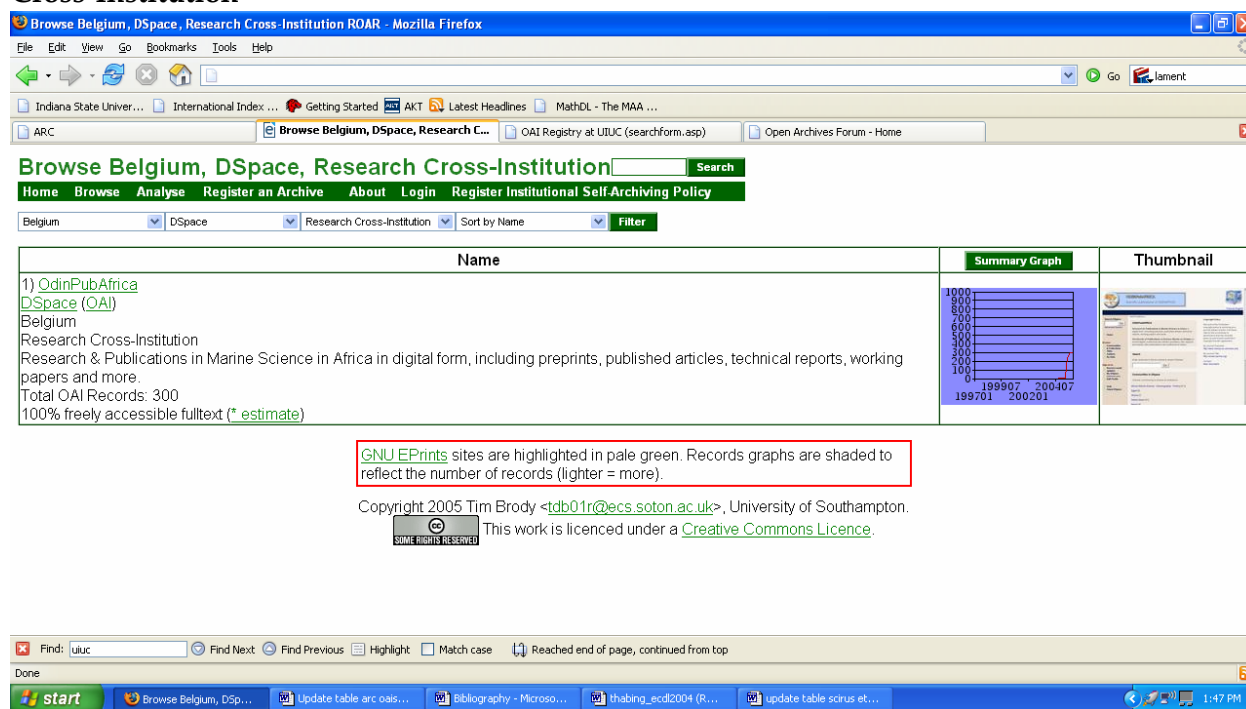
Source: http://archives.eprints.org/ (February 28, 2006)

The Browse feature gives composite record counts by three major parameters: country, archive type and software. Record counts are limited to those archives registered and successfully harvested by Celestial; the figures are not restricted to full-text items but reflect all metadata records.

Table 09: ROAR Statistics of Archive Type

ARCHIVE TYPE	Archives	In Celestial	Records	Mean	Median
Research Institutional or Departmental	314	248	757,286	3,054	272
e-Journal/Publication	66	43	172,905	4,021	120
Research Cross-Institution	63	54	1,792,048	33,186	569
e-Theses	63	52	333,097	6,406	674
Demonstration	24	12	5,533	461	28
Database	11	5	2,056	411	160
Other	94	60	601,345	10,022	176
TOTAL	**635**	**474**	**3,664,270**	**57,561**	**1,999**

Source: http://archives.eprints.org/index.php?action=browse (February 28, 2006)

ROAR's categorization of "archive types" is unique. Given its focus on e-prints, it is not surprising to find that "research institutional or departmental" deployments account for nearly half of ROAR's archives. There is little doubt that this broad category also subsumes some e-journal/publication and e-theses content. These three categories combined account for 70 percent of the archives but only 37 percent of the records, whereas "research cross-institution" accounts for less than 10 percent of the archives but nearly 50 percent of the records. The record count could be quite different if all the archives were fully represented in Celestial or if the archives in the "other" category (94) were assigned to a discrete category. [29]

Table 10: ROAR Statistics of System Software Deployments

SYSTEM SOFTWARE (# of deployments, if readily available from software Web site)[30]	Archives	In Celestial	Records	Mean	Median
GNU EPrints (UK) (198) http://www.eprints.org/software/archives/	196	176	120,513	685	164
DSpace (USA) (136) http://wiki.dspace.org/DspaceInstances/	131	82	175,227	2,137	403
Bepress [Digital Commons] (44) http://www.umi.com/proquest/digitalcommons/	43	25	58,178	2,327	504
ETD-db (USA) http://scholar.lib.vt.edu/ETD-db/	22	18	263,364	14,631	1,295
OPUS: Open Publications System (Germany) (39) http://elib.unistuttgart.de/opus/doku/about.php?la=en	21	18	5,073	282	79
DiVA (Sweden) (15) http://www.diva-portal.org/about.xsql	14	13	8,966	690	387
CDSware (Switzerland)	8	5	103,201	20,640	3,339

[29] On review, it appears that many of them do, in fact, adhere more appropriately to an existing category.

[30] Software URLs, country of origin, and number of instantiations added by the author (March 12, 2006).

http://cdsware.cern.ch/cdsware/overview.html					
ARNO (Netherlands) (6)					
http://arno.uvt.nl/~arno/site/	5	4	171,402	42,851	16,801
DoKS: Document & Knowledge Sharing (Belgium)					
http://doks.khk.be/wiki/index.php/Main_Page	3	3	2,170	723	226
HAL: Hyper articles en Ligne (France)					
http://hal.ccsd.cnrs.fr/index.php	3	3	52,650	17,550	1,089
Fedora (USA) (32)					
http://www.fedora.info/community/	2	2	208	104	104
eDoc (Greece)					
http://www.edocplus.com/company/overview.htm	2	2	39,770	19,885	19,885
MyCoRE (Germany)					
http://www.mycore.de/	1	1	1,935	1,935	1,935
Other software (various)	184	122	2,661,613	21,817	595
TOTAL	**635**	**474**	**3,664,270**	**146,257**	**46,806**

Source: http://archives.eprints.org/index.php?action=browse (February 28, 2006)

ROAR offers easy access to information about which archives utilize specified system software. Almost all the archives deploying a handful of major repository software systems are fully represented in ROAR (e.g. GNU EPrints, DSpace, DiVA, ARNO, Digital Commons bepress). Although there are myriad IR systems in use worldwide, it would be helpful if more of the archives falling into the "other software" were reviewed and either placed into an existing or newly-created software category (e.g. Arc; Archimède; digitAlexandria—FreeScience and Archivemaker; DLXS; and the Public Knowledge Project's Open Journals and Open Conference Systems). At present the "other category" represents 29 percent of archives and a whopping 73 percent of ROAR's records. Among the top twenty largest archives in ROAR, thirteen presently fall into the "other software" category (e.g., CiteSeer, PubMed Central, arXiv, Library of Congress's American Memory).

As advocates of self-archiving and the Open Access principles set forth in BOAI and the Berlin Declaration, ROAR also operates a registry of institutional self-archiving policies, recently renamed ROARMAP (Registry of Open Access Repository Material Archiving Policies). As of mid-May 2006, 19 institutions in nine countries and one European-wide research institution had registered a policy commitment. Each entry includes a link to the institutional repository, its growth data, and details about its OA policy. Five institutions mandate self-archiving: CERN, University of Southampton, Queensland University of Technology, University of Minho, and University of Zurich. ROARMAP includes model self-archiving policy statements; model policies for national and private research funding agencies are also presented.

In May 2006, ROAR announced two new developments. First, in addition to the RSS, plain-text and ListFriends exports, its records (not their content) became OAI-compliant, initially available as Dublin Core. Secondly, as part of the Preserv project (http://preserv.eprints.org) they added support for Content Profiling institutional

repositories—available for most GNU EPrints and DSpace repositories. Users can access links from ROAR entries to the Preserv Profile link for those repositories with functioning (and registered) OAI interfaces. This generates a graph showing the breakdown of all file formats contained in the repository. Users can click on a format's red bar to obtain a complete listing of identified records.

Figure 09: Screenshot of Preserv Profile, Lincoln University, Faculty of Technology repository

Source: http://archives.eprint.org (May 13, 2006)

4.1.5 OpenDOAR: Directory of Open Access Repositories

Launched to the public in late January 2006 by the University of Nottingham and University of Lund (developer of DOAJ), OpenDOAR is sponsored by the Open Society Institute (OSI), the UK's Joint Information Systems Committee (JISC), the Consortium of Research Libraries (CURL, British Isles), and SPARC Europe. Created to support the Open Access movement, OpenDOAR aims to categorize and build a "comprehensive and authoritative list" of OA research archives worldwide.[31] Ultimately, the directory will "serve not only as a discovery tool for scholars seeking original research papers or specific digital representations, but also as a developmental tool for repository administrators and service providers who want to build new services tailored to targeted user communities" (Hubbard 2005).

---

[31] http://www.doar.org/

OpenDOAR staff verify the data about each repository, "noting new features and directions," in order to enrich and enhance future versions of the directory service. The repositories listed in OpenDOAR have been surveyed by researchers as opposed to automatically identified and listed. This approach is valuable (although initially resource-intensive) when compared to some auto-harvested listings, according to its proponents, because roughly 40 percent of repositories surveyed have been rejected as out-of-scope or non-functional.

As of May 2006, the directory lists 380 repositories and offers repository-level keyword searching or browsing with filters by country, content type or subject. Eventually OpenDOAR expects to classify repositories by other parameters and also offer the capacity to search within repositories. Results can be presented in full or short format.

Table 11: OpenDOAR Statistics of Content Type and Subjects (February 2006)[32]

Content Type		% of Total N=353	Subjects		% of Total N=353
Articles	218	61.8%	Agriculture and Food Sciences	63	17.8%
Books	110	31.2%	Arts and Architecture	113	32.0%
Chapters	98	27.8%	Biology and Life Sciences	147	41.6%
Conference papers	146	41.4%	Business and Economics	149	42.2%
Dissertations	212	60.1%	Chemistry	116	32.9%
Learning objects	27	7.6%	Earth and Environmental Sciences	139	39.4%
Multimedia	28	7.9%	Health Sciences	134	38.0%
Patents	7	2.0%	History and Archaeology	113	32.0%
Posters	23	6.5%	Languages and Literatures	133	37.7%
Pre-print journal articles	89	25.2%	Law and Political Science	142	40.2%
Presentations	33	9.3%	Mathematics and Statistics	148	41.9%
Reports	148	41.9%	Philosophy and Religion	110	31.2%
Research datasets	3	0.8%	Physics and Astronomy	124	35.1%
Software	6	1.7%	Science General	82	23.2%
Undergraduate theses	72	20.4%	Social Sciences	234	66.3%
Working papers	66	18.7%	Technology and Engineering	218	61.8%

Source: http://www.opendoar.org/ (February 28, 2006)

Unlike ROAR, categorizations are not mutually exclusive. According to OpenDOAR's data, the vast majority of repositories represent a mix of content types (an average of 3.6 different types of materials per repository) and subjects (an average of six different subjects per repository). The utility of the present categories is questionable due to their scope and redundant use. Articles, dissertations, reports and conference papers

---

[32] In mid-May 2006, OpenDOAR covers 380 (as opposed to 353) repositories.

dominate the content, with very few repositories registering datasets, software or patents. In terms of subject categories, the Social Sciences content surpasses all categories (perhaps reflecting redundancy with the Business and Economics, and Law and Political Science categories); Technology and Engineering is a close second. Most subject categories are quite evenly distributed (falling in the distribution range of 32 to 42 percent).

Aligning OpenDOAR's typologies with the repository descriptions is problematic and it is hard to imagine how a system that requires a high level of OpenDOAR staff intermediation will scale up. For example, an institutional repository of a research organization in France (ALADIN) working in the "humanities and social sciences" that "will include articles, technical reports, working papers, images, videos, and more," is coded by two subjects—Earth and Environmental Sciences, and Social Sciences—and by three content types—Articles, Working Papers and Reports. This narrower categorization evidently reflects OpenDOAR's initial focus on research papers and related materials (e.g., theses); expansion of content type listings is desired and intended, given continued funding for this initiative.

Figure 10: OpenDOAR sample search result for Social Sciences

ALADIN: Accès Libre aux Archives du Dépôt Institutionnel Numérique de la MSH-Alpes
Country: France
Organization: La Maison des Sciences de l'Homme-Alpes
Subjects: Earth and Environmental Sciences --- Social Sciences
Type: Articles --- Working papers --- Reports
OAI Base URL: http://dspace.msh-alpes.prd.fr/oai/
Description: ALADIN is a pilot project for publications produced by researchers and partners of MSH-Alpes. MSH-Alpes is a public basic-research organization working in the scientific field of humanities and social sciences (and depending upon CNRS and different Grenoble universities). Ultimately this repository will include articles, technical reports, working papers, images, videos, and more.

Source: http://www.opendoar.org/ (March 2006)

The Aristotle University of Thessealoniki Document Server in Greece "contains theses, articles, papers and photos" and is coded as Articles, Dissertations, and Multimedia and with four broad subject codes. This falls far short of characterizing the repository's content or alerting users to its collections (e.g., historical collection of Greek newspapers-1800 to present, photographic archive of traditional 18th-20th century art, or archaeological events in Greek press-1832 to1932). Nor is the repository retrieved when a user searches for Greek newspapers, newspapers Greece, newspapers or archaeology. Since there is only one repository from Greece, it can be retrieved by country.

Despite these shortcomings, OpenDOAR is in its early stages of deployment and aims eventually to serve multiple user groups "each with their own expectations, needs and perspectives" making it possible to search, filter, analyze and query the descriptions of

each repository in customizable and meaningful ways. Closer collaboration—or an eventual merger—with ROAR seems desirable and would allow combining the best features of each service, as informed by user feedback.

4.1.6 Arc: Cross Archive Search Service

Update Table 02: Arc based on DLF Survey responses, Fall 2005

	Arc http://arc.cs.odu.edu/
ORGANIZATIONAL MODEL	Old Dominion University w/out base funding
SUBJECT	Multi-disciplinary
FUNCTION	Cross archive digital search service that harvests OAI-compliant repositories.
PRIMARY AUDIENCE	Research community; Digital library developers
STATUS	Experimental research service
SIZE	7,156,192 records (64% increase) from 177 archives (8.5% increase)
USE	No response
ACCOMPLISHMENTS	1. Maintained since 2003. 2. Successful experimentation on Lucene indexing to replace database indexing. 3. Successful experimentation on distributed storage on PC-cluster. 4. Arc open source software in SourceForge is used by other projects inside and outside ODU.
CHALLENGES	1. Performance problems in grouping search results by archives, subjects, etc. 2. Large volume of data requires fundamental change of architecture. 3. Incremental complexity of source code calls for addressing extensibility.
TOOLS OR RESOURCES NEEDED	1. Apache Struts framework to restructure into multi-layered MVC pattern. 2. Apache Lucene indexing framework to speed up the metadata searching and retrieving.
GOALS OF NEXT GENERATION RESOURCE	1. Deployment of Lucene/cluster version. 2. Investigate how to provide richer service by integrating Web2.0 technology.

As was the case in 2003, users are informed that Arc "is an experimental research service of Digital Library Research group at Old Dominion University. Arc is used to investigate issues in harvesting OAI compliant repositories and making them accessible through a unified search interface. It is not a production service and may be subject to unscheduled service interruptions and anomalies." In fact, Arc was unstable during the five-month period while this report was written, making it difficult to evaluate fully. Arc

researchers report that they have been working on a fast, parallel search-based, robust new version that should be available by mid-June 2006. It is based on Lucene parallel indexing.

Arc contains more than seven million metadata records, including 4.3 million from OCLC's XTCat (bibliographic records of dissertations and theses extracted from WorldCat, which has been static since its initial harvest several years ago). During the six-month period of this review, Arc remained static in size. Access to the "Administration" page that contained details about the last harvests when this service was reviewed in 2003 is now restricted and inaccessible.

With few exceptions, Arc's search and retrieval functions have not changed since the last report was released nor have the problems identified in conducting searches been addressed (further evidence that Arc is intended for R&D purposes—not for end-users). However, two new features are worth noting for their (as yet unrealized) potential usefulness. In advanced search mode, there is an option to "search the last results" or conduct a "new search." In addition, queries can be limited within a specified archive to particular "archive sets." In most instances, unfortunately, the archive only has the default option—"all sets"—available; however, two examples with "archive sets" illustrate the value of this feature. A search of the University of Nottingham's repository can be limited to one of eight constituent departmental archives; similarly the National Science Digital Library (NSDL) development site at Cornell can be filtered to eight different collections. In cases where repositories have meaningful sub-collections of materials, this filtering device would prove very useful.

In "Lessons learned with Arc, an OAI-PMH Service Provider," Liu et al. (2005) inform readers how Arc—which introduced the concept of "hierarchical harvesting" that formed the basis for OAI aggregators—has served as the platform for other projects including Archon (described in the 2003 DLF report and included in Appendix 3 of the current report), Kepler (enables self-archiving by means of an "archivelet"), the Networked Computer Science Technical Reference Library (NCSTRL), and DP9 (an OAI gateway service for Web crawlers). Among more recent initiatives undertaken by the Department of Computer Science at Old Dominion University, the Digital Library Grid, funded by The Andrew W. Mellon Foundation, is developing software tools that take advantage of grid computing so that costs associated with federating heterogeneous digital libraries are more effectively distributed, thereby improving sustainability. "Because of Arc's immense scale," these researchers rightfully conclude, "it has informed the community on a number of issues related to synchronization, scheduling, caching, and replication." Their current work will "merge OAI-PHM digital libraries with grid computing," helping to secure the technical architecture and infrastructure required by large-scale operations (Liu et al. 2005, 602).

4.1.7 OAIster

Update Table 03: OAIster based on DLF Survey responses, Fall 2005

	OAIster http://www.oaister.org/
ORGANIZATIONAL MODEL	U of Michigan w/ initial Mellon funding; now IMLS in collaboration w/ DLF, UIUC and Emory. New Yahoo! Search and Google partnerships.
SUBJECT	Multi-disciplinary
FUNCTION	Collection of freely available, difficult-to-access, academically-oriented digital resources which are easily searchable.
PRIMARY AUDIENCE	Academic community
STATUS	Established
SIZE	6,000,000 records (300% increase) from 550 institutions (182% increase)
USE	Per month: 15K to 19K hits. 100s of 1,000s via Yahoo! Search
ACCOMPLISHMENTS	1. Increase in size and use. 2. Development of OAI Best Practices. 3. Respect for OAI (e.g., most vendors incorporate it now). 4. Modifications to advanced search and inclusion of Book Bag feature.
CHALLENGES	1. Changes in departmental focus may reduce OAIster priority. 2. Need to recruit programmer.
TOOLS OR RESOURCES NEEDED	1. UTF-8 tools that permit harvester to verify if record is UTF-8 or not and communicate that effectively, with appropriate display, to data providers. 2. Streamlined method for maintenance and indexing.
GOALS OF NEXT GENERATION RESOURCE	1. Z39.50/MetaLib integration (accomplished as of spring 2006). 2. Clustering analysis for better search and browse. 3. Better informed through user feedback. 4. Many interface and functionality tweaks.

With Arc serving primarily as a research test-bed, OAIster is the only large-scale OAI multidisciplinary aggregator operating as a full production service for the benefit of end-users. OAIster harvests metadata on a weekly basis and prominently notes new "institutions" and new record counts on its home page. (This was recommended in the 2003 DLF report.) Growing by leaps and bounds, as of mid-May 2006, OAIster harvested five times the number of metadata records from more than triple the number of institutions as it did in mid-2003.

A hallmark of OAIster is that it limits harvesting to OAI-compliant records that have full digital representation associated with the item (e.g., full text, digital image, etc.); however, it is important to note that OAIster's definition of "freely available" includes

some full-text licensed resources. The most prominent example is the inclusion of the Institute of Physics' journal articles (210,000 records), but there are others such as African Journals Online (18,000 records). OAIster is currently re-thinking its collection parameters with the intent of broadening its scope to embrace items with restricted access to full-text. In addition to providing users with a collection development policy, it would be helpful if OAIster's search results marked items only accessible through licenses or if it permitted users to filter results by restricted versus non-restricted access.

Since 2003, three enhancements to OAIster's user interface stand out. First, "dataset" was added as a "resource type," making it possible to limit searches to this medium. A keyword search for "data" coupled with the filter to retrieve "datasets" returned 280,495 results. Searches can be refined or limited by selecting among the institutions highlighted in the left-hand frame. Twenty-one institutions hold datasets, and at a glance, it is evident that the vast majority of them (279,286) come from one source—PANGAEA: Publishing Network for Geoscientific and Environmental Data. The second enhancement dates from November 2005 when OAIster deployed a "bookbag" feature, enabling users to save records during a session and download or email them. Most recently, in March 2006, OAIster added "language" as a search field option. A search for <Afrikaans> returns one dissertation from the Netherlands but <German> returns more than 74,000 results. (More than half of these records are from Bibliotheksservice-Zentrum Baden-Wurttemberg although more than 120 different archives in OAIster hold German-language materials.)

OAIster makes a vast reservoir of digital content available, but constructing effective searches is not always straightforward, requiring, for example, an understanding of how terms are combined and nested. As evident from the following search results for dissertations on global warming, the first two terms are nested together and then coupled with the third term:

> Global warming AND thesis OR dissertation
> > Retrieves 112,373 items
> > Interpreted as (global warming and thesis) or dissertation; thus retrieving any item tagged as a dissertation irrespective of the subject.

> Thesis OR dissertation AND global warming
> > Retrieves 87 items
> > Interpreted as (thesis or dissertation) and global warming; thus retrieving either theses or dissertations about global warming.

Many entries are lengthy; users would benefit from the option to select short or full displays. The search query will also return items using the word "thesis" when it refers to an argument or proposition. If the data provider includes dissertation or thesis in the resource-type field, OAIster would normalize the metadata and these records could be

retrieved by limiting the search by "text." If a record does not include those terms, of course, they will not be discoverable. OAIster's clustering effort (described below) aims to support more granular resource-type options via a drop-down menu (including "dissertation" and "thesis"). Admittedly, this is only a partial solution since OAIster must rely on what information the metadata record includes.

Many enhancements depend on the concerted efforts of data providers, achieved by conforming to accepted standards and best practices. For example, effective date searching hinges on more widespread uniformity in the metadata expressing dates. When asking, "why normalize," OAIster's Kat Hagedorn illustrates the wide variance in expressing dates in OAIster:

WHY NORMALIZE?

Sample date values in OAIster:

<date>2-12-01</date>
<date>2002-01-01</date>
<date>0000-00-00</date>
<date>1822</date>
<date>between 1827 and 1833</date>
<date>18--?</date>
<date>November 13, 1947</date>
<date>SEP 1958</date>
<date>235 bce</date>
<date>Summer, 1948</date> (Hagedorn 2005b).

OAIster is exploring how to adapt CDL's date normalization utility to help overcome these inconsistencies.[33]

Browsing by topical categories relies on appropriate metadata subject tags from data providers. And searching within institutions/collections depends on archives providing "sets" that reflect meaningful sub-collections. For these reasons OAIster's developers are among the key proponents of improving and enriching metadata through DLF's best practices. OAIster is also experimenting with visualization and semantic clustering techniques based on work at Emory University (e.g., MetaCombine project, see SouthComb in section 4.4.9),[34] UIUC (e.g., refer to the prototype CIC Metadata Portal in section 4.1.8), and UC-Irvine.

---

[33] CDL Date Normalization Utility (DNU) Documentation, Landis and Loy, last updated August 24, 2005, available from
http://www.cdlib.org/inside/diglib/datenorm/datenorm_documentation.doc.
[34] MetaCombine Project (Emory University) is available from http://www.metacombine.org

Among the more vexing problems, not only for OAIster but affecting other aggregators as well, is managing duplicate records. As Khan et al. (2005) attest, duplication is easy to eradicate when two records have identical metadata fields, but difficult to detect when they differ slightly (for example, due to data entry errors or different practices in expressing an author's name). Using a subset of data from Arc as a test bed, the authors demonstrate a duplication detection algorithm they developed which might be applied to other large aggregations like OAIster.

OAIster has identified the improvements that it intends to make as time permits:

- Show HTML embedded in records. Make HTML embedded in search results records viewable and linkable.
- More relevancy sorting options. Potential to order results by proximity, institution frequency, among other options.
- Date searching. Single date and date range searching.
- Searching within institutions. Choice of institutions to search in.
- Browsing capability. Browsing of broad topical categories of records.
- Duplicate records. Handling of records that are the same among repositories.

Bugs to be fixed:

- Highlight words or phrases in results list when punctuation exists.
- Count resource type search hits in hit frequency and weighted hit frequency sorts.
- Correct secondary sorting for date ascending and date descending sorts. (Source: http://oaister.umdl.umich.edu/o/oaister/future.html.)

OAIster was among the first OAI data providers to collaborate with Yahoo! Search and Google; OAIster sends them metadata on a monthly basis. Yahoo! Search uses the complete metadata records in their search index, whereas Google uses the URLs included in the records to find pages for their search index. These partnerships facilitate deeper indexing than available via regular Web crawling.[35]

In March 2006, OAIster announced the availability of its metadata for use by federated search engines via SRU and created a Web page with instructions about how it use its metadata outside OAIster's interface (http://oaister.umdl.umich.edu/o/oaister/sru.html).

---

[35] For information about Yahoo! Search refer to http://search.yahoo.com/. Peter Suber has written a guide, "How to facilitate Google crawling: Notes for open-access repository maintainers," available from http://www.earlham.edu/~peters/fos/googlecrawling.htm .

External referrals from general search engines may account for 20 or more times the number of queries than direct OAIster searches.[36] While precise data is scarce on the topic, ProQuest has analyzed Web traffic to its Digital Commons' repositories and reports that most users (95 percent) find their way to OAI content via general search engines. This trend decreases slightly over time as users become aware of the repository: after the first year of deployment, external referrals dropped to 75 percent. A growing number of institutional repositories, such as the University of Minho (Portugal), are starting to make OAIster directly searchable from their sites as illustrated by the screenshot below.[37]

Figure 11: Screenshot of the Universidade do Minho (Portugal) IR with OAIster Integration
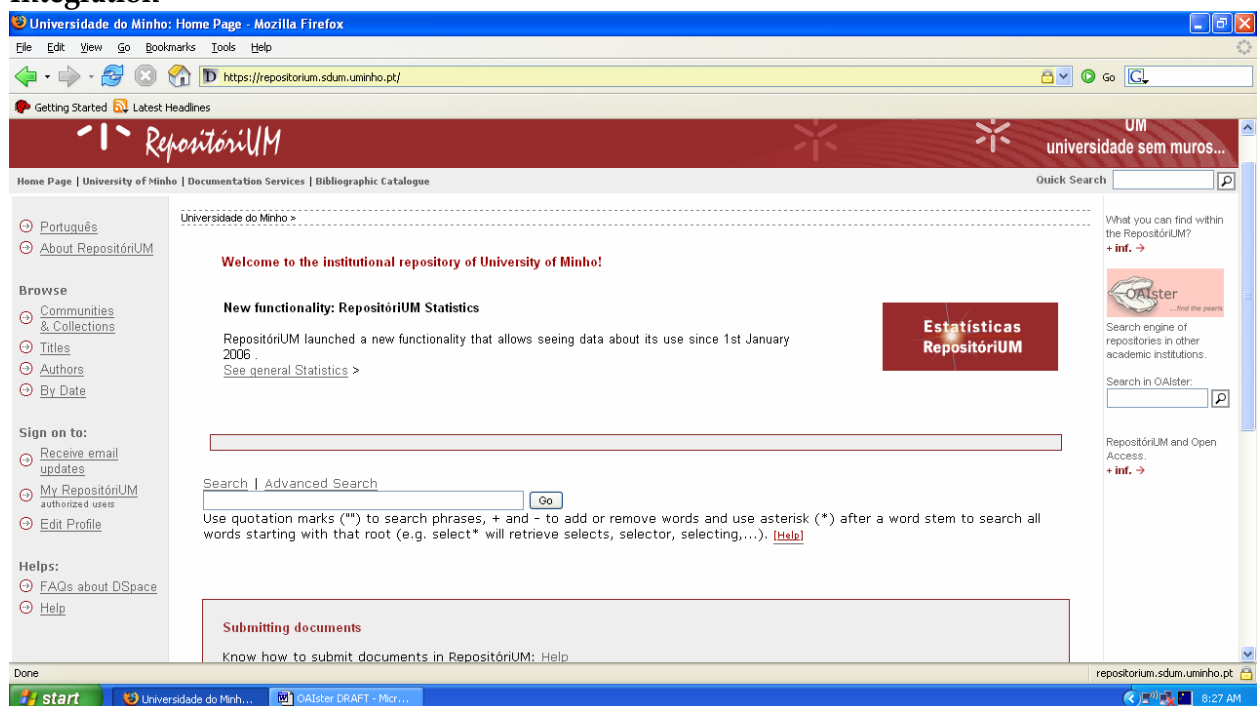
Source: https://repositorium.sdum.uminho.pt/ (May 12, 2006)

The prominent inclusion of OAIster helps researchers see how their work fits into a larger scholarly communication framework, bringing increased visibility and the potential for wider impact. For instructions to replicate this integration, refer to "Using OAIster Metadata Outside this Interface" available from OAIster's home page.

---

[36] Email correspondence with Jeff Riedel on December 8, 2005 and phone interview on December 13, 2005.
[37] The University of Minho is one of few institutions to mandate self-archiving: http://www.eprints.org/openaccess/policysignup/.

4.1.8 Consortial Portals: CIC Metadata Portal, DLF Portal, DLF MODS Portal

Update Table 04: CIC Metadata Portal and DLF Portals based on DLF Survey responses, Fall 2005

CIC Metadata Portal http://cicharvest.grainger.uiuc.edu/ http://nergal.grainger.uiuc.edu/cgi/b/bib/oaister	DLF OAI Portal http://www.hti.umich.edu/iml/	DLF MODS Portal http://www.hti.umich.edu/m/mods/
ORGANIZATIONAL MODEL		
Collaboration with CIC member libraries.	DLF members and allies with OAI records.	DLF members and allies who publish OAI records that contain MODS metadata as well as the basic Dublin Core record.
SUBJECT		
Cross-disciplinary	Predominantly humanities and cultural heritage	Cultural heritage
FUNCTION		
Research issues relating to consortial metadata aggregation describing both freely available and restricted license content.	To publicize publicly-accessible holdings of DLF member institutions.	A testbed to demonstrate the value of MODS records in the provision of richer library services.
PRIMARY AUDIENCE		
Academic Community	Academic Community	Academic Community
STATUS		
Under Development	Under Development	Experimental
SIZE		
517,000 records from 171 academic collections from 10 CIC universities	883,992 records from 44 repositories	253,478 records from four repositories (Indiana, LC, OCLC, U of Chicago)
USE		
Not available	Not available	Not available
ACCOMPLISHMENTS		
1. Incorporation of rich collection descriptions into the search. 2. Generation of thumbnail images included with search results. 3. Incorporation of data from harvested resources (not just OAI) into search indexes. 4. Normalizing & enhancing metadata to support various browse & search interfaces.	1. Creating it. 2. Growing it. 3. Using it to solicit feedback from scholars on ways to improve.	1. Launching it. 2. Modifying it after meeting with DLF Scholars Advisory Panel in June 2005. 3. Added thumbnails; bookbag feature; improved sorting for date, title and author; simple vs. advanced searching modes.

CHALLENGES		
1. Resources to maintain the service.	1. Local OAI skills. 2. Willingness to make harvestable metadata a local priority.	1. Getting feedback from users. 2. Getting libraries to publish MODS records.
TOOLS OR RESOURCES NEEDED		
Money	Training (which DLF is providing).	Programmers.
GOALS OF NEXT GENERATION RESOURCE		
Uncertain since it is a research project and not a production service.	Roll it out to the public. Grow it aggressively, both in bulk and quality.	To continue to prototype services as articulated by DLF user community and DLF Aquifer.

CIC Metadata Portal

Founded in 1958, the CIC is an academic consortium of the eleven institutional members of the Big Ten Athletic Conference plus the University of Illinois at Chicago and the University of Chicago. The CIC Metadata Portal is a collaborative pilot project undertaken to research issues related to aggregating metadata and testing different user interfaces. As of December 2005, the CIC metadata repository contained more than 550,000 records harvested from 187 digital collections held by eleven of the thirteen CIC member institutions. Nearly half of the records (267,000) are contributed by the University of Michigan; the University of Illinois at Urbana-Champaign accounts for another 22 percent (~125,000). Participating institutions adopt the general CIC collection policy and metadata guidelines. Resources include a wide spectrum of types of information. An estimated 70 percent of the records refer to digital objects (have a referring URL); an estimated 50 percent are restricted access, only available to those universities with licenses to access the content.[38]

The portal uses the University of Michigan DLXS software also deployed by OAIster, and therefore, exhibits similar advanced search functionality including searching by field, filtering by resource type, and user-control over the ways in which results are sorted. The CIC portal has several resource types not available via OAIster that allow users to limit their queries to sheet music, theses, software and Web sites (but not datasets). It also utilizes an automated process to generate thumbnails and thumbshots from the URLs pointed to in the metadata records (Foulonneau, Habing and Cole 2006). Thumbnails are provided at both the collection and item-level. As of December 2005 only an estimated 35,000 item-level records had thumbnails.[39]

---

[38] Muriel Foulonneau, CIC metadata portal: Project status, Powerpoint presentation at the Big Ten Center, Chicago, December 12, 2005.

[39] Information about the automated generation of thumbnails is also available at the project Web site: http://cicharvest.grainger.uiuc.edu/thumb.asp.

Figure 12: Screenshot of CIC Metadata Portal Search Page

Source: http://nergal.grainger.uiuc.edu/cgi/b/bib/oaister (April 30,2006)

From the CIC search portal, users can conduct simple searches, view "featured collections," or browse collection-level records by institution. Unlike OAIster and the DLF portal (described below), the CIC portal has not deployed a Book Bag function that permits users to save results within a session.

The CIC is experimenting with four innovative user interfaces:

- **Faceted access** permits "who, what, when, and where" searches.
  http://nergal.grainger.uiuc.edu/cgi/b/bib/oaister?page=newpage
- **Geographic browse** offers map-based discovery and display of results—of special interest because the resources cover an estimated 175 countries and 80 languages.
  (Password-protected while under development.)
- **Collections browse** links to collection-level descriptions with thumbnails (where available).
  http://cicharvest.grainger.uiuc.edu/colls/collections.asp
- **EAD (Encoded Archival Description) test portal** containing metadata from institutions with EAD finding aids.
  http://nergal.grainger.uiuc.edu/cgi/f/findaid/findaid-idx

Although the CIC Metadata Portal is not a production service, it has furthered research about effective collaboration and produced a number of promising applications (Foulonneau et al. 2006).

DLF OAI Portal

The DLF OAI Portal, in an early stage of development as of May 2006, is a metadata repository containing more than one million items from 45 DLF collections/institutions. DLF's membership includes major research libraries in the United States that are leading the way in digital library innovation, along with a small but influential number of international partners. As a result, this aggregation contains some of the finest digital collections, coming from such prestigious institutions as the Library of Congress, the California Digital Library, Cornell University, Emory University, the University of Chicago, the University of Illinois, Urbana-Champaign, and the universities of Indiana, Michigan, Pennsylvania and Virginia. Once fully developed with more complete holdings from repositories at the Bibliotheca Alexandrina, the British Library, Columbia, Harvard, New York Public Library, Princeton, Stanford and Yale, this portal will offer access to a rich aggregation of premier digital collections.[40]

Utilizing the DLXS software, the user interface has the unadorned look and feel of OAIster. It supports simple and advanced searches (Boolean operators applied to keyword, title, author/creator/, subject, and language) as well as delimiters by resource type (text, image, audio, video, and dataset).

As is the case with OAIster, "Browse Institutions" represents a mélange of both high-level composite general collection descriptions (e.g., Indiana University's Digital Library's multiple digital collections are represented by a composite entry) and specific digital collections within an institution (e.g., the University of Pennsylvania is represented as several "institutions" with separate entries for various digital projects). The descriptions represent both the specificity of information provided by the institution as well as the number of separate data repositories deployed within an institution. In short, there is one description in "Browse Institutions" for each repository in the portal. Users, however, would benefit from a more uniform representation of what constitutes a "collection." After updating its contents, the DLF Collections Registry (described in section 4.4.3) and the DLF OAI Portal need to harmonize their collection/institution descriptions.[41] The figures below show the difference in the way Indiana University is represented in the DLF OAI Portal (and OAIster), the DLF Collections Registry, and the CIC Metadata Repository. The user is at a loss to know how many "collections" IU's digital library hosts: three, eight or seventeen?

---

[40] A list of DLF partners and affiliates is available from http://www.diglib.org/about.htm.
[41] See DLF Digital Collections Registry, browse by Institution, available from http://gita.grainger.uiuc.edu/dlfcollectionsregistry/browse/GEMHostInst.asp

Figure 13a: Indiana University's digital collections (3 of them in bold typeface) as described by the DLF OAI Portal (and OAIster)

Indiana University Digital Library Program (26857 records)
http://dlib.indiana.edu/

The Indiana University Digital Library Program is dedicated to the selection, production, and maintenance of a wide range of high quality networked resources for scholars and students at Indiana University and elsewhere. The program provides OAI-enabled access to the **U.S. Steel Gary Works Photograph Collection, 1906-1971**, the **Frank M. Hohenberger Collection** and the **Sam DeVincent Collection of American Sheet Music**.

Figure 13b: Screenshot of Indiana University's digital collections (17 of them) from the DLF Collections Registry
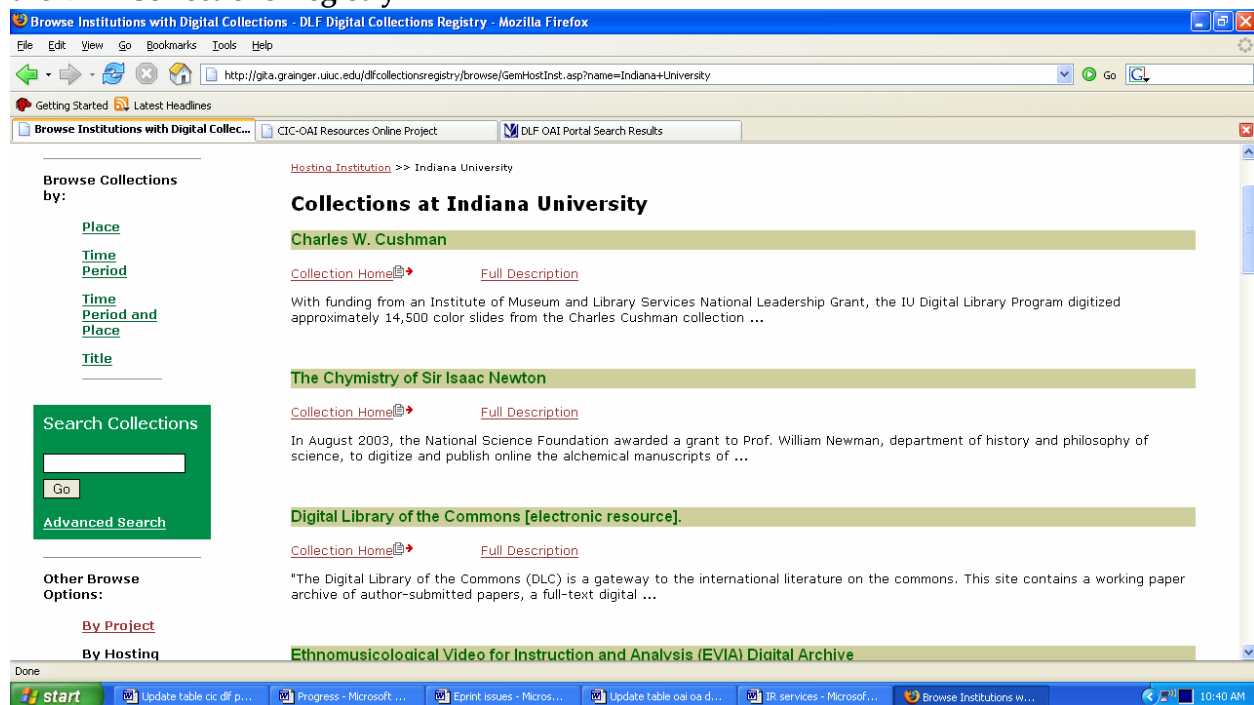
Figure 14: Indiana University's digital collections (8 of them) as represented in the CIC Metadata Portal
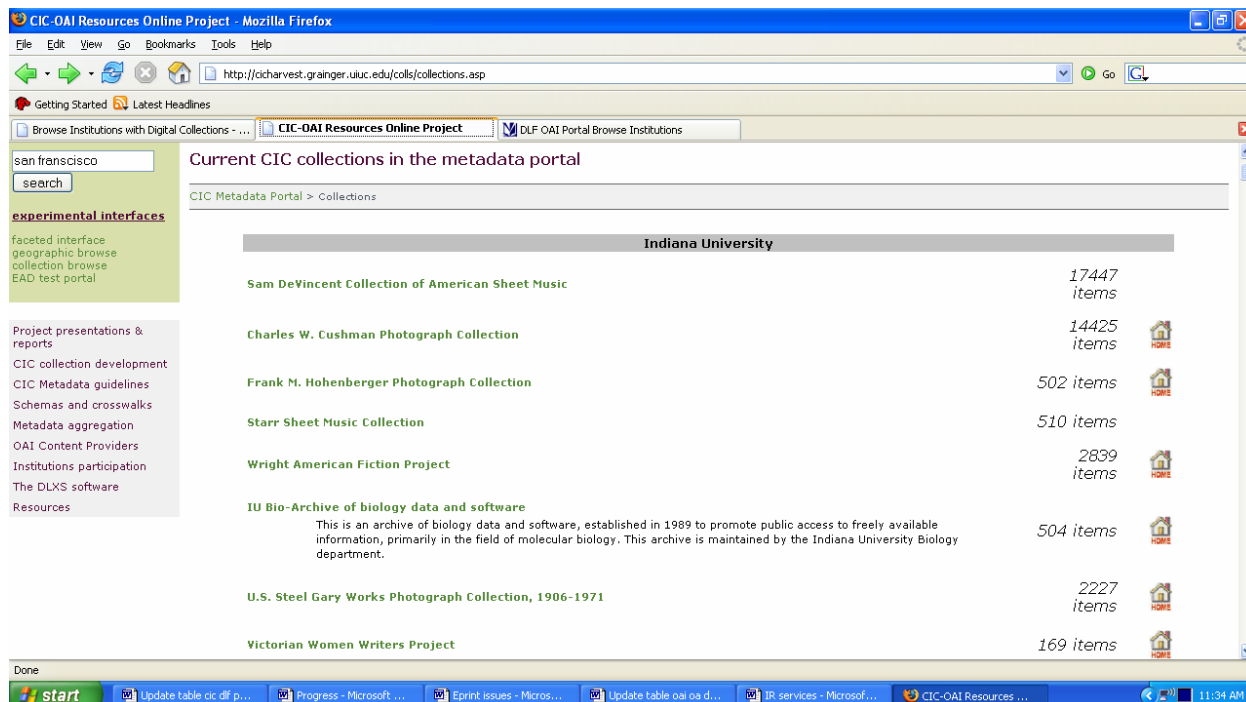
Sources: http://www.hti.umich.edu/i/imls/viewcolls.html; http://gita.grainger.uiuc.edu/dlfcollectionsregistry/browse/GemHostInst.asp?name=Indiana+University and http://cicharvest.grainger.uiuc.edu/colls/collections.asp (April 30, 2006).

DLF MODS Portal

The DLF MODS Portal, developed with funding from the DLF's current IMLS grant, is the testing ground for new features and functionality that are subsequently ported to OAIster. Among its accomplishments (noted in the description of OAIster above as well) are the inclusion of thumbnails, the bookbag feature, user-choice of simple or advanced searching modes, and improved capabilities for sorting results by date, title and author.

In an early stage of deployment, the DLF MODS Portal also serves as a prototype to test out the enriched Metadata Object Description Schema. The MODS element set is richer than Dublin Core but simpler than full MARC. As of mid-May 2006, this portal contains more than 250,000 MODS records from four institutions:

- Indiana University Digital Library Program (certain sets)
- Library of Congress Digitized Historical Collections
- OCLC Research Publications
- University of Chicago Library Metadata Repository (certain sets)

Figure 15: Screenshot of Record from DLF OAI Portal
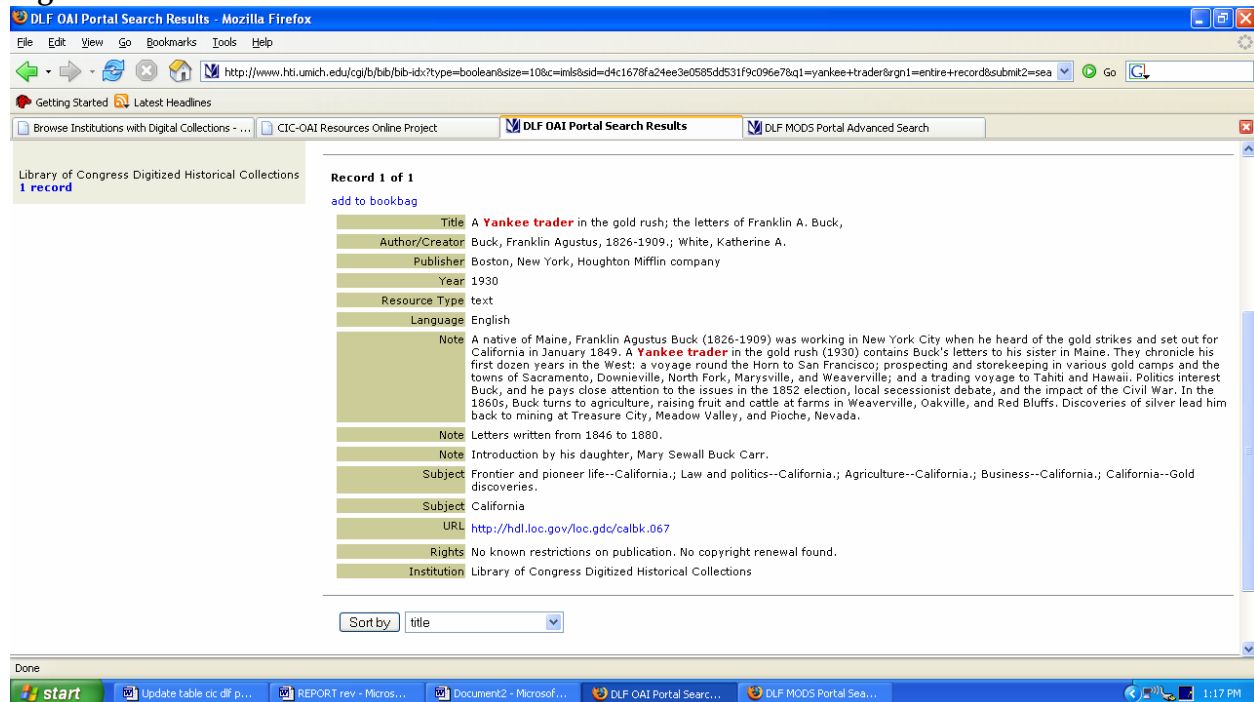
Source: www.hti.umich.edu/cgi/b/bib/bib-idx?c=imls;page=simple

Figure 16: Screenshot of Richer Record from DLF MODS Portal
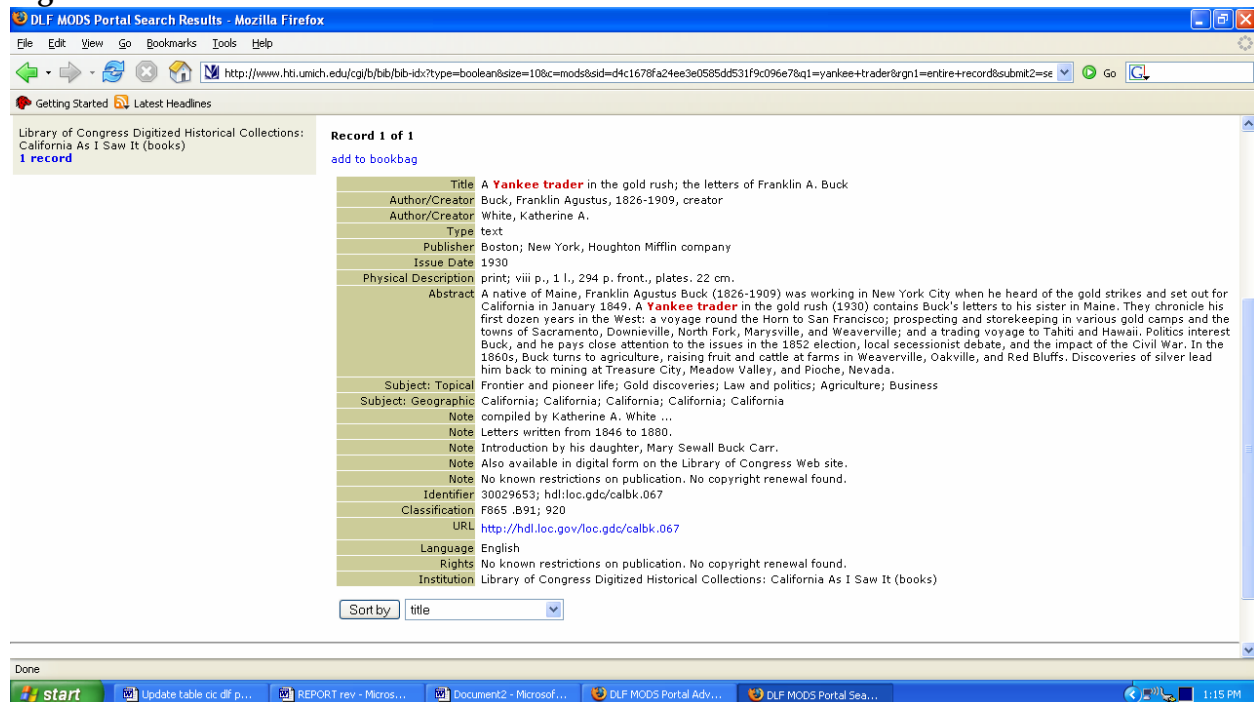
Source: http://www.hti.umich.edu/m/mods/ (April 2006)

The screenshots above show the differences in record display for two different metadata implementations for the same object, *A Yankee Trader in the Gold Rush; The Letters of*

Franklin A. Buck, from the Library of Congress American Memory collection. This comparison between the DLF OAI and DLF MODS portal reveals how the enriched MODS record, with its more specific tagged fields, makes possible enhanced search and retrieval functions.

The DLF Aquifer project (see section 4.4.8) will also require contributing institutions to use the MODS standard for bibliographic data. The DLF MODS Portal will continue to evolve based on needs of the DLF user community and the DLF Aquifer Project.

4.1.9   Germany: OA and OAI Access Points

DINI (Deutsche Initiative für Netzwerkinformation E.V.) in Germany exemplifies a coordinated national approach to OA and OAI adoption. In addition to organizing workshops to promote the Open Access and self-archiving, DINI maintains a centralized directory of OA repositories, establishes quality control through a repository certification process, and operates an OAIster-like search engine across German OA repositories. The directory can be searched or sorted by place, university, URL, contact person, OAI interface, and DINI certification.

> The DINI certificate distinguishes the repository from common institutional web servers and assures potential users and authors of digital documents that a certain level of quality in repository operation is warranted. In addition, DINI sees its certificate as an instrument to support the Open Access concept. (Dobratz and Schoger 2005)

A separate search engine, DINI OAI Search Engine (OAI-suche) for German Open Access Repositories, currently conducts searches across 50 German libraries, archives and document servers, comprising 44,336 items. Repositories are harvested on a weekly basis and statistics about the number of records and most recent harvest dates are readily available. Content is searchable by author, title, keyword, or abstract and queries can be limited by language, date, date range or archive. Users can pre-select whether results should be returned by date and they can control the number of returns per page. A search for <wirtschaft> (economics) returns 740 results with briefly annotated entries and links to full-text content.

The Electronic Journals Library (EZB--Elektronische Zeitschriftenbibliothek), with nearly 31,000 titles (an estimated 12 percent are e-only), is arguably the world's largest database of scholarly electronic journals. Operated by the University of Regensburg, EZB represents a consortium of 343 libraries that pool bibliographic information and metadata about freely available and licensed e-journals subscriptions. Ninety-four percent of all German university libraries (n=77) participate along with 80 percent of German national and central subject libraries (e.g., constituents of the Max-Planck

Institute). Full-text accessibility is indicated by color-coded dots. An estimated 41 percent of all titles are freely available in full text (i.e. Green).

Figure 17: Dot color-coding scheme in EZB

🟢	Full texts are freely accessible.
🟡	The library / research institute has a license for this journal; therefore it is accessible for the users of this institution.
🔴	The journal is not on subscription, thus full texts are not accessible. Mostly, however, tables of contents and in many cases abstracts are available free-of-charge.
🟡🔴	The institution has no continuous subscription on this journal. Therefore, only some of the published volumes are accessible as full texts.

Source: http://rzblx1.uni-regensburg.de/ezeit/about.phtml

Journals are browsable by forty-one different subject areas or by title. Nine subject areas have 400 or more "green" titles (or 63.6 percent of the freely available full-text e-journals).

Table 12: EZB Subjects with 400 or more "green" titles

	# of Titles	% Free Full-text
Medicine	5,525	33.1%
Economics	2,706	39.9%
Biology	1,587	28.5%
Political Science	1,403	55.3%
Sociology	1,067	39.9%
History	1,027	61.0%
Law	1,056	55.7%
Agriculture & Forestry	880	48.8%
Education	751	55.5%

Source: Based on data from the Electronic Journals Library: Annual Report 2005 (April 2006).

In contrast, Chemistry & Pharmacy is represented by eleven hundred titles but only 20 percent are freely available in full-text (221 titles).

Users can search for journals by various fields including title, keyword and publisher with the option to limit queries to specific subjects. Through the "preferences" Web page, users can select particular regions or institutions and conduct searches to display their holdings. EZB partnered with the German subject gateway, Vascoda, to incorporate e-journal titles into discipline-specific virtual libraries.[42]

---

[42] Information about the Vascoda project available from http://www.bibliothek.uni-regensburg.de/projekte/vascoda/vascoda.htm.

Figure 18: Screenshot of EZB Titles in the Anglo-American Culture & History Virtual Library
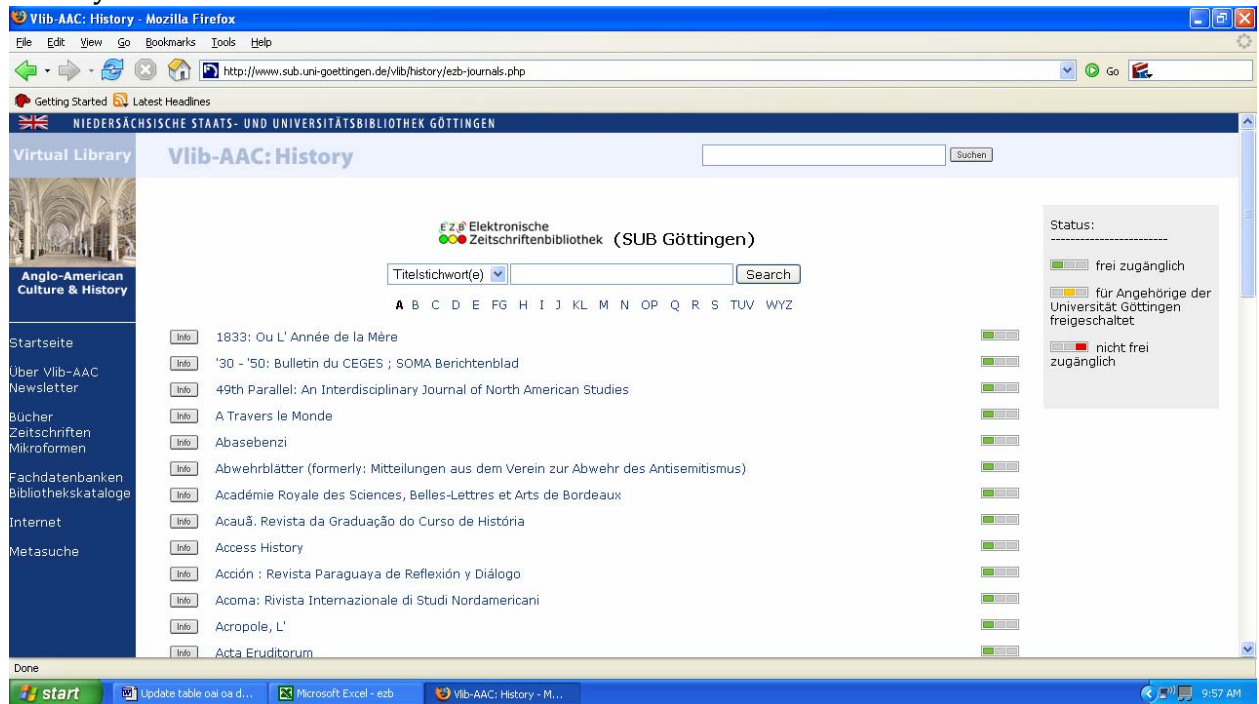
Source: http://www.sub.uni-goettingen.de/vlib/history/ezb-journals.php (March 24, 2006)

More than forty information services incorporate EZB's content through OpenURL linking. Currently EZB is working with Vascoda to streamline authentication and permissions so only a single sign-on is required to access licensed resources.[43]

4.1.10   Current Issues and Future Directions

- These services now contain a wealth of information. In general, they warrant more widespread marketing and use. At the same time, it would be beneficial to better understand the characteristics of their users and the nature of their uses.
- "Open access" and "freely available" may carry different meanings in these services. Users are not as concerned about the fine points of definitions, but they would like to know the scope of coverage, what is or is not included. Items that are restricted to licensed users should be clearly indicated.
- In many instances it is difficult to distinguish records representing metadata-only from those that also link to full-object representation. Users may wish to have access to the broader spectrum of resources, but should be able to decipher whether or not additional content is available and under what circumstances.
- Application of visualization and clustering tools (by subject, geographic area, time period) helps users to interpret and navigate through large results sets.

---

[43] More information about this project is available from http://aar.vascoda.de/

- The database management information from many of these resources is of great value to analyzing the growth in digital repositories worldwide. This data should be readily available for mining by any interested user, ranging from journalists to academics.
- The synergistic relations between these services help to foster enhanced OAI-compliance, improved coverage, broader use of resources, and better communication between OAI data and service providers. Examples include cooperative efforts between DOAJ and OpenDOAR, OpenDOAR and ROAR, and the UI OAI Registry and OAIster. Further collaboration might lead to more uniform agreement of terminology and better delineation of service coverage while reducing redundancy (e.g., multiple technical registries for OAI-PMH and overlapping lists of publisher/journal self-archiving policies)
- A recent comparative study (the first of its kind) that investigated coverage of the "OAI-PMH corpus" by three general search engines found that Yahoo indexed 65 percent, followed by Google with 44 percent, and MSN with 7 percent (McCown et al. 2005). According to the researchers, 21 percent of the resources were not indexed by any of the three search engines. The authors suggest that if these popular search engines supported OAI-PMH directly, it would increase interest in registering and implementing OAI-PMH repositories. They conclude: "Search engines would benefit by being able to index more content, and DLs would benefit by being able to share their contents with search engines without incurring web crawling overhead."
- It might prove worthwhile to call a summit of the core OAI registries and general OAI search services to discuss how to better market their services, not only by extending the reach of their content into these generic popular search engines but also by attracting more users directly to their sites. This would build on various options already deployed such as RSS feeds, A9.com open search, Firefox search engine plug-in, and the development of OA toolbars like OASes, geared to academic users.[44]

---

[44] The Center for History and New Media, George Mason University, is building a scholar's Web browser, Firefox Scholar (aka SmartFox) with a late summer 2006 beta release http://echo.gmu.edu/toolcenter-wiki/index.php?title=Firefox_Scholar_%28aka_SmartFox%29. OASes is a toolbar for Internet Explorer designed for students and scholars that searches six OA resources (OAIster, DOAJ, PubMed Central, Creative Commons, Project Gutenburg, and FindArticles) and four general search engines (Google, Yahoo, MSN, and Clusty). Created by Dr. Shahul Ameen (M.D. and Senior Resident at the Central Institute of Psychiatry, Ranchi, India), it also offers a pull-down menu of links to Open Access News, BOAI, SPARC, PLoS, and NASA ADS. OASes is free to download and use without registration. Available at http://www.psyplexus.com/oases/.

4.2 Links in the Scholarly Communication Value Chain

Changes in the landscape of scholarly communication over the past few years come into sharp focus through a review of how e-print services are evolving. As discussed earlier in this report, in the short span of time since the original report appeared, the open access movement has gained international momentum and engendered a multitude of commitments from major funding agencies, intergovernmental organizations, private and public foundations, university and library consortia, publishers and single institutions.[45] Stemming in large part from self-archiving and harvesting of research output from e-print repositories, the aggregations described in this section represent various subject-based services, along with affiliated discovery and citation analysis tools. Connected together, they serve vital functions in the scholarly communication value chain supporting registration, certification, awareness, archiving and rewarding of intellectual capital (see figure 19, Van de Sompel et al. 2004).

The specific services reviewed here include four varieties of self-archiving and aggregating content: discipline-driven, centralized, author self-archiving of preprints (arXiv); research agency-driven, centralized archiving of technical reports and harvesting of related archives (NASA Technical Reports Server and CERN Document Server); semi-mandated author or publisher centralized self-archiving of peer-reviewed journal articles (PubMed Central); and community-driven centralized deposit of domain-based literature (Open Language Archives Community). Each of these services was also reviewed in the 2003 DLF survey; the discussion here updates and expands on the earlier report.

Special consideration is given to electronic theses and dissertations (ETDs) because they represent a prevalent form of research output. Often aggregated in repositories at the institutional level, ETDs also form the basis of an international community of practice via the Networked Digital Library of Theses and Dissertations. Recent activities to coordinate ETD deployment at the national and transnational level in Europe are described. Finally, tools for discovering ETDs are discussed, most notably Elsevier's Scirus ETD search engine.

The University of Illinois's Grainger Engineering Library OAI Aggregation serves as a cross-repository niche search engine, harvesting records from more than 50 data providers including other services discussed in this report (e.g., arXiv, CDS, DOAJ, NSDL). Covering similar territory, PerX, a pilot search engine developed in the UK for engineering, is briefly described. Future DLF studies should include discussion of the

---

[45] Refer to Peter Suber's "Timeline of the Open Access Movement" for a chronology of milestones, available from: http://www.earlham.edu/~peters/fos/timeline.htm and to his annual summaries and predictions in *SPARC Open Access Newsletter* (January issues 2003 to present), available from: http://www.earlham.edu/~peters/fos/newsletter/archive.htm .

U.S., Department of Energy, Office of Scientific & Technical Information (OSTI) E-Print Network Search service (http://eprints.osti.gov/).[46]

CiteSeer and Citebase round out this section and represent services that support reference linking and citation analysis of research literature. CiteSeer focuses on computer science, aggregating literature via Web crawling and data mining techniques in addition to supporting self-archiving, whereas Citebase covers a broader subject domain in the sciences through OAI harvesting. It is beyond the scope of this report to examine recent parallel services such as Google Scholar (http://scholar.google.com/), Microsoft Academic Search (http://academic.live.com/), and Thomson Scientific's Web Citation Index (http://scientific.thomson.com/free/essays/selectionofmaterial/wci-selection/), but it is important to note that they draw their inspiration and to varying degrees, their core technology, from CiteSeer.

4.2.1 arXiv

Update Table 05: arXiv based on DLF Survey responses, Fall 2005

	arXiv http://www.arxiv.org
ORGANIZATIONAL MODEL	Originally LANL, now Cornell with partial NSF support.
SUBJECT	Science: physics, math, non-linear science, computer science, quantitative biology
FUNCTION	Automated e-print archive server; rapid distribution system prior to peer review.
PRIMARY AUDIENCE	Research community
STATUS	Established
SIZE	340,000 articles (nearly 50% increase)
USE	Per year: 16.8 million unique full-text downloads per year; Per month: 4,000 submission
ACCOMPLISHMENTS	1. Creation of quantitative biology section. 2. Established user endorsement system. 3. New interface for computer science section (CoRR).
CHALLENGES	1. Continuous heavy use. 2. Staff time & funding. 3. Integration of legacy features/code with new developments.
TOOLS OR RESOURCES NEEDED	Money and time.
GOALS OF NEXT GENERATION RESOURCE	1. Reduced admin time through better facilities. 2. Easier submission process for users. 3. Additional features: flexible alerting, dynamic classification, etc. 4. Better integration with other scholarly resources.

---

[46] OAIster began to harvest OSTI metadata in March 2006. As of late May 2006, OSTI's OAI repository had more than 125,000 items.

At fourteen years old, arXiv.org remains the earliest, largest and most successful example of a subject-based e-print archive, with readership and monthly submissions growing steadily. Warner reflects on "lessons learned" and charts arXiv's evolution from a "self-contained preprint redistribution service" to a key component of "an integrated global communication system" (2005, 58). ArXiv's content is integrated into federated searches and harvested by aggregators on a worldwide basis.

ArXiv was conceived as a means to formally communicate and rapidly disseminate research progress, not to replace peer-reviewed journals which are considered indispensable to certification and reward systems. Indeed, arXiv has served as a nexus of innovation by demonstrating "how conventional peer review can be implemented on top of an open access substrate," for example, through the creation of journals such as *Advances in Theoretical and Mathematical Physics, Geometry and Topology, Logical Methods in Computer Science* and all journals of the Institute of Mathematical Statistics (Warner, 2005, 58-59). Both the American Physical Society and the Institute of Physics (UK) accept direct electronic article submissions from arXiv.

Warner discusses the importance of "community" (through the creation of subject advisory boards) and "critical mass" to arXiv's success. To ensure high quality, relevant submissions, in January 2004, arXiv instituted an "endorsement system" that requires most new users to receive ratification from another user prior to submitting their first paper. To support this endorsement system and provide authors with a list of papers they have written, arXiv has established "authority records" that link a person's arXiv account with their papers.

In terms of rights and permissions, Warner explains that for many years "arXiv operated without any explicit statements about rights"; it was assumed that the act of submission granted arXiv the non-exclusive right to distribute the paper. Several years ago, arXiv instituted a license click-through as part of the submission process in which the author:

- grants arXiv.org a license to distribute this article;
- certifies the right to grant this license;
- understands that submissions cannot be completely removed once accepted; and
- understands that arXiv.org reserves the right to reclassify or reject any submission. (Warner 2005, 64)

Currently other options are under consideration—either simply granting arXiv a license to distribute or agreeing that a Creative Commons license applies, which provides the requisite permissions.

ArXiv created a proxy submission site in France as part of HAL (hypertext articles online at Center for Direct Scientific Communication in Lyon) whereby submissions in

relevant subject categories are automatically transferred to arXiv (unless the depositor expressly prohibits it). Similarly, documents for which the full text is already available in arXiv (or TEL—French Theses online) do not need to be uploaded again into HAL; the insertion of a link in HAL makes the file visible.[47]

Using arXiv as the exemplar, in "Rethinking Scholarly Communication," Van de Sompel et al. (2004) postulate about new ways to combine the five functions of scholarly communication:

- *Registration*, which allows claims of precedence for a scholarly finding.
- *Certification,* which establishes the validity of a registered scholarly claim.
- *Awareness,* which allows actors in the scholarly system to remain aware of new claims and findings.
- *Archiving,* which preserves the scholarly record over time.
- *Rewarding,* which rewards actors for their performance in the communication system based on metrics derived from that system. (Van de Sompel et al. 2004, citing the work of Roosendaal and Geurts 1997)

They depict the information flow of an e-print from its entry point in arXiv through "multiple services hubs that fulfill functions of the scholarly communication process." The authors illustrate how multiple players and pathways interact in the value chain of scholarly communication (Figure 19). Disciplinary archives, like arXiv may serve four of five functions, while services like Citebase (see section 4.2.10) discharge some of the reward functions through the provision of citation metrics.

---

[47] Information available from HAL's Welcome page: http://hal.ccsd.cnrs.fr/index.php?langue=en

Figure 19: arXiv ecology and the emergence of service pathways [48]
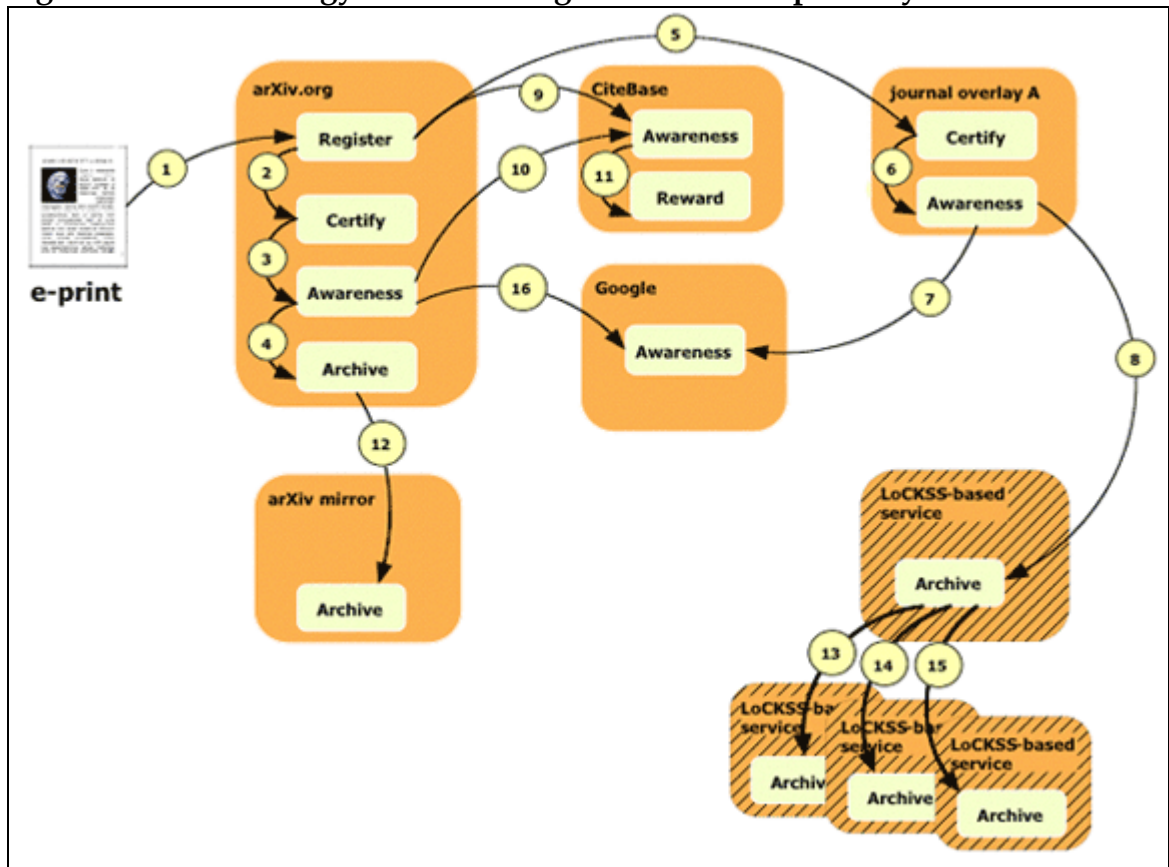
Reproduced with permission of the authors.

When looking to the future, Warner suggests that it is too early to determine what impact institutional repositories will have on arXiv, speculating that the "intermediate stage will be for arXiv to act as a slave subject-based publishing venue with institutional repositories serving as the primary archives, or vice versa" (2005, 67). In the long term, the funding model of institutional repositories, which is more closely aligned with its direct beneficiaries, may prove more viable than arXiv's situation, where the Cornell community comprises only a minor constituency among arXiv's global authors and readers, but has fiduciary responsibility for operating the service with NSF contributing some research funding.

---

[48] "Each step in the information flow is shown as a numbered arrow." (Van de Sompel et al. 2004)

4.2.2 NTRS: NASA Technical Reports Server[49]

Update Table 06: NTRS based on DLF Survey responses, Fall 2005

	NASA Technical Reports Server (NTRS) ntrs.nasa.gov
ORGANIZATIONAL MODEL	NASA
SUBJECT	Science: aerospace and other related scientific areas
FUNCTION	Technical Report Server to collect, archive and disseminate scientific paper.
PRIMARY AUDIENCE	Research Community; Interested Public
STATUS	Established
SIZE	902,000 records (63% increase) of which ~495,000 full-text (~125,000 from NASA agencies; most not free).
USE	Per day: 17K unique daily visits. Per month: 30,000 full-text downloads.[50]
ACCOMPLISHMENTS	1. Improved OAI tools (e.g., OAI GW to harvest data from master archive at NASA Center for AeroSpace Information). 2. Improved user interface. 3. Normalized data.
CHALLENGES	1. Integrating video. 2. Integrating natural language query capabilities. 3. Indexing full text.
TOOLS OR RESOURCES NEEDED	COTS applications to meet challenges and requirements.
GOALS OF NEXT GENERATION RESOURCE	1. Better user interface. 2. Improved data mining capabilities.

The NASA Technical Report Server (NTRS) aggregates more than 900,000 metadata records from 18 agencies, 40 percent of which are derived from four external (non-NASA) services. Among the fourteen NASA agencies covered, the Center for AeroSpace Information (CASI) is by far the largest, contributing some 540,000 metadata records about 23 percent of which represent full-text documents. The significant growth in content aggregated by the NTRS is due primarily to an increase in records from CASI, the Jet Propulsion Laboratory (not covered in 2003), and the Department of Energy, Office of Scientific and Technical Information's "Information Bridge" (OSTI). Not only have CASI's metadata records nearly doubled but its full-text documents have grown from 100 to more than 90,000. Although according to its Web site, "NASA citations and full-text documents found on NTRS are unlimited, unclassified, and publicly available," most full-text technical reports are not free-of-charge, but can be ordered from NASA.

---

[49] A listing of more than 9,000 NASA-related acronyms is available from the NASA STI home page (http://www.sti.nasa.gov/STI-public-homepage.html: see "Tools/Products/Services," or directly from http://www.sti.nasa.gov/acronym/main.html).

[50] Nelson et al. (2004, 1).

Since the 2003 DLF survey, NTRS use has increased dramatically from an estimated 6,500 searches per month to 17,000 unique visits daily in late 2005.

Over the past two years, resources from one NASA agency have been removed due to unresolved copyright issues, the Goddard Institute for Space Studies,[51] and another added, the Dryden Flight Research Center (589 full-text papers). As evident from Table 13, five other NASA agency sites are static; NTRS has not recorded any harvests or updates since July 2004. Correspondence with NASA officials reveals that the records for four of the agencies (GENESIS, Goddard, Kennedy and Stennis) were obtained by isolated Web crawls and that RIACS (Research Institute for Advanced Computer Science) has ceased operation of its e-prints software system.[52] (RIACS technical reports can be downloaded directly from its Web site.)

Table 13: NTRS Constituent Archives

NASA ARCHIVES[53]	# Records 2006	# Records 2003	% Full text	Downloads of full text 4/28/03 to 6/30/04	Download rank = # of documents N = 312,115	Most recent harvest or update Status on 2/7/2006
GENESIS (NASA Jet Propulsion Laboratory)	37	27	100%	403	11	**2/3/2006**
NASA Ames Research Center	354	354	0%	52[54]	14	7/9/2004
NASA Center for AeroSpace Information (CASI)	507,371	256,637	23%[55]	1,269	8	**12/6/2005**
NASA Dryden Flight Research	589	N/A	100%	N/A	N/A	**2/3/2006**

---

[51] More than 2,000 GISS research reports are available from http://pubs.giss.nasa.gov/

[52] Email correspondence on February 10, 2006 with Michael L. Nelson, who is no longer under contract with NASA but oversaw these Web crawls in 2004.

[53] The sources for the data are as follows: # of records 2006, gathered from the NTRS "About the Collections" on February 8, 2006; # of records 2003, from Brogan (2003, 21); % full text, # of downloads and download rank, from Nelson and Bollen (2005, 393); most recent harvest or update from NTRS "About the Collections" on February 8, 2006. In respect to NASA Arc which has 354 metadata records none of which are full text, the authors explain that at the time of their study, it had some content that has since been removed.

[54] At the time the study was conducted NASA Arc (and several other agencies) had a little bit of full content according to email correspondence from Michael L. Nelson on February 10, 2006. Some full-text publications were identified through Web crawls, but these reports were removed as a precaution when NASA personnel discovered that not all appropriate document availability authorization (DAA) forms were on record.

[55] At the time the study was conducted CASI had about 5% full content. According to email correspondence with Calvin E. Mackey on February 8, 2006, it now has 90,507 full-text documents (or 23 percent).

Center						
NASA Goddard Space Flight						
Center	11	11	100%	1	17	7/9/2004
NASA Jet Propulsion Laboratory	19,570	N/A	100%	65,508	3	**2/3/2006**
NASA Johnson Space Center	129	128	80%	2,413	6	**2/3/2006**
NASA Kennedy Space Center	82	82	100%	2	16	7/9/2004
NASA Langley Research Center	5,090	3,948	100%	151,524	1	**2/3/2006**
NASA Marshall Space Flight						
Center	571	498	100%	4,493	5	**2/3/2006**
NASA Stennis Space Center	39	39	100%	14	15	7/9/2004
National Advisory Committee for						
Aeronautics (NACA)	7,640	7,639	100%	72,122	2	**2/3/2006**
NIX Images	0	N/A	N/A	N/A	N/A	**1/18/2006**
RIACS (NASA Ames Research						
Center)	0	61	100%	390	12	7/2/2004
NASA Goddard Institute for Space						
Studies (GISS)	[1,771]	1,335	40%	809	10	N/A
Subtotal NASA Agencies	**541,483**	**270,759**		**299,000**		
NON-NASA ARCHIVES						
Aeronautical Research Council						
(UK)	2,647	2,647	100%	10,184	4	**2/3/2006**
arXiv Physics Eprint Server	272,266	243,707	100%	1,181	9	7/10/2004
BioMed Central	18,454	17,507	100%	166	13	7/9/2004
Information Bridge: Energy						
Citations Database (OSTI)	76,473	20,738	70%	1,584	7	**2/3/2006**
Subtotal Non-NASA Archives	**369,840**	**284,599**		**13,115**		
GRAND TOTAL	**911,323**	**555,358**		**312,115**		

Among the four external archives, only two are actively harvested—the UK's Arc service, which comprises historical documents (and is also static at 2,647 reports)[56], and OSTI, which continues to grow. Neither arXiv nor BioMed Central, despite their continual growth, have been harvested or updated at the NTRS site since July 2004.

---

[56] The Aeronautical Research Council (Arc), the principal agency in the UK producing technical reports on aeronautics, existed from 1909-1979 and published reports until 1980. The AERADE (aerospace and defense) Reports Archive at Cranfield University (UK) offers unified searching of the 10,000 historical digitized reports from Arc and NACA. (The National Advisory Committee for Aeronautics (NACA) was chartered in 1915 and operated as a precursor to NASA from 1917 to 1958.) The digitized report collection is available from http://aerade.cranfield.ac.uk/reports.html. AERADE reports that there are many more Arc documents in its collection but lack of funding has prevented their digitization. Currently they offer a scan-on-demand service with a charge of $40. For information about reports available in printed format, consult the Library Catalog available from http://unicorn.central.cranfield.ac.uk/uhtbin/webcat/. Information based on email correspondence with John Harrington on February 10, 2006.

Harvesting of these two services was possibly curtailed (users are not informed) as a result of NASA's emphasis on upgrading the functionality of their own publications and the technical capabilities of the contractor operating NTRS.[57] This narrowing of focus is supported by an examination of user log files from April 2003 to June 2004 that shed data[58] on which NTRS repositories received the most downloads. "While contributing significantly to the total number of holdings in NTRS," Nelson and Bollen found that the "Energy Citation Database [OSTI], BioMed Central and arXiv.org contributed little to the download totals" (2005, 393). The authors postulate that the prominent number of downloads from NACA and the UK's Arc "suggests an interest in historical aeronautical publications."[59] They also speculate that users are most interested in aerospace-focused materials and that the "presence of other STM [scientific, technical, medical] materials has yet to expand its user base." Noting that arXiv is harvested by a host of other services, Nelson and Bollen conclude that its presence does "not guarantee its use in NTRS" (Nelson and Bollen 2005, 393).

Search Features

Whereas Simple Search defaults to NASA-only agencies, in Advanced Search users are given the option to select among twelve NASA agencies and four external archives. If a deliberate decision was made to cease from actively harvesting metadata from arXiv and BioMed Central, users are not warned from either Advanced Search (which is used twice as much as simple search according to Nelson and Bollen) or from the "Help" page. Users need to consult "About the Collections," browse by archive and sort results by date added to NTRS, or utilize the "Weekly Update" function to ascertain the status of harvests and updates for each service.

According to NTRS's News Archive, searches were expanded in September 2004 to include accession and document identification numbers. In July 2005, NASA's Scientific and Technical Information Program Office announced the implementation of persistent unique identifiers for all public full-text documents (NASA 2005).[60]

New User Interface

In February 2006, a new public interface for NTRS will launch, featuring direct searching of text files and searching within a browse function (or vice versa) (NASA 2006). According to the January 2006 pre-launch announcement, users will be guided by navigation menus that are recalculated with each new search. When large result sets are

---

[57] Hypotheses put forward by Michael L. Nelson in email correspondence on February 10, 2006.
[58] The author gratefully acknowledges Brian Lavoie et al. for the concept of "shedding data" (Lavoie et al. 2006).
[59] See footnote above regarding unified access to UK Arc and US NACA historical documents.
[60] Information about the Digital Object Identifier System is available from www.doi.org . About Handle Systems refer to http://www.handle.net/ .

retrieved, customized refinement options are presented to the user. Customized browsing options will enable users to look for new related information. The new system also offers automatic spelling corrections and "did you mean…?" suggestions. "Navigation and search options are captured in the browser URL," permitting users "to save and share any view of data by bookmarking the link or cutting and pasting it into an email message." Search results are relevance-ranked and sortable. The new NASA interface utilizes the Endeca Guided Navigation search engine.

The recommendation service (linking from the results page to recommended related documents) instituted by NTRS in September 2003 was terminated, although this is not noted in the News Archive.[61] However, NASA officials are quick to point out that Phase 2 of the new interface (anticipated in summer 2006) will have "recommendation like services." Among its features:

> 1. The system can automatically retrieve top/most requested items. The same data and rules-based decisions can also be used to display the top articles in a particular area or could be combined with application logic to retrieve the most requested items in that area.

> 2. The system can show related items to the user as they navigate through the result set. This system allows the organization to define rules that also show related information as the user browses through result sets.

> For example, the most popular authors for the current result set could be listed along side the main result set.

> These rules can be prioritized and only the most relevant items will be shown to the user. This dynamic rules-based retrieval of additional information can be applied to the entire site or only specified areas of the site.[62]

Phase 2 will also incorporate multimedia.

NTRS as Hierarchical Aggregator

As an OAI hierarchical aggregator, NTRS offers the potential advantage of convenient, one-stop shopping for other OAI service providers (Nelson et al. 2003). The scientific

---

[61] Refer to the user study about the recommendation service (Nelson et al. 2004). According to follow-up email correspondence with Michael L. Nelson on February 10, 2006, the early results of the recommendation service were encouraging and their article outlined plans to improve it (e.g., by eliminating the "cold start" problem,) but the contractor responsible for running NTRS lacked the technical capacity to maintain it. Nelson confirmed that the recommendation service has been terminated.

[62] Email correspondence from Calvin E. Mackay on February 10, 2006.

search engine, Scirus harvests four NTRS collections (GENESIS, Langley, Marshall and NACA), totaling 12,265 full-text records; OAIster harvests seven (Ames Research Center, CASI, Goddard, Kennedy, Langley, Stennis, and the UK's Arc), totaling 3,466 records; and NSDL eight (GENESIS, Ames, Goddard, Johnson, Kennedy, Langley, Stennis and UK Arc), totaling 8,288 records. OAIster harvests directly from five NASA agencies rather than relying on the NTRS aggregation (e.g., GENESIS, Dryden, Johnson, Marshall, and NACA). (OAIster does not harvest from any collections that do not point to freely available digital objects, e.g., full-text documents). Representatives from NSDL report that NTRS sets are complex and problematic, returning many failed messages. Although NSDL would like to cover more NTRS resources, since mid-December 2005, its only successful NTRS harvest is UK Arc metadata.[63]

4.2.3 PubMed Central

Update Table 07: PubMed Central based on DLF Survey responses, Fall 2005

	PubMed Central http://www.pubmedcentral.nih.gov/
ORGANIZATIONAL MODEL	U. S. National Library of Medicine
SUBJECT	Science: life sciences
FUNCTION	Voluntary publisher-based archiving of live sciences journal literature.
PRIMARY AUDIENCE	Research Community
STATUS	Established
SIZE	430,000 articles (330% increase) from 200 journals (54% increase); ~6,000 new articles deposited per week.
USE	Per month: 960,000 unique IP addresses (unique users est. at 1.5 times that number); 2.8 million full-text articles retrieved, > 6 million pages retrieved.
ACCOMPLISHMENTS	1. OAI service is operational. 2. More than 250,000 retrospective scanned articles. 3. UK's Wellcome Trust digitization collaboration. 4. NLM Journal Article DTD gaining wide acceptance.
CHALLENGES	No response
TOOLS OR RESOURCES NEEDED	No response
GOALS OF NEXT GENERATION RESOURCE	No response

Since launching its OAI service in October 2003, PubMed Central (PMC), the National Institutes of Health's (NIH) free digital archive of full-text life sciences journal literature and data managed by the National Library of Medicine (NLM), has become the third

---

[63] NSDL's experiences in harvesting OAI metadata data from heterogeneous sources are described more fully in Section 4.3.1 and chronicled by Lagoze et al. (2006a).

largest resource in OAIster (after Picture Australia and CiteSeer). It ranks first, in the category of OAI-compliant peer-reviewed, full-text journal article aggregations; second only to HighWire Press in the number of freely available articles. (HighWire Press, which is not a fully OAI-compliant service, boasts nearly 1.2 million free, full-text articles from 918 journals.) PubMed Central has quadrupled in size over the past two years, providing access to 430,000 articles (including more than 250,000 retrospective scanned articles) from 200 journals by fall 2005. With the advent of its OAI service, PMC also began to accept individual open access articles from journals, such as *Science* and *Biological Chemistry* that are not regular contributors to PMC.

In May 2005 the NIH put into effect a public access policy, specifying PMC as the central repository of articles emanating from NIH-funded research. According to the policy, researchers are requested to submit to PMC the final version of their peer-reviewed electronic manuscript no later than twelve months after its publication in a scientific journal. NIH offers three primary reasons for endorsing public access:

- Archive - A central archive of NIH-funded research publications preserves these vital published research findings for years to come.
- Advance Science - The repository is an information resource for scientists to mine more easily medical research publications and for NIH to manage better its entire research investment.
- Access - The policy provides patients, families, health professionals, scientists, teachers, and others electronic access to research publications resulting from NIH-funded research. [NIH Public Access, Policy Overview available from http://publicaccess.nih.gov/overview.htm][64]

PMC was chosen as the central repository because it is publicly accessible, a permanent archive, and searchable.

In the early implementation of NIH's new public access program (NIHPA), submissions are estimated below four percent of the eligible articles.[65] Upon review of these low deposit statistics, the NIH Public Access Working Group recommended a policy change to require deposit by researchers. The CURES Act, introduced before Congress in December 2005 includes a provision supporting public access to federally-funded medical research. ARL reports that "under the proposed legislation, articles published in a peer-reviewed journal would be required to be made publicly available within 6

---

[64] FAQ from NIH about its public access policy is available from: http://publicaccess.nih.gov/publicaccess_QandA.htm

[65] According to Peter Suber's calculations NIH grants resulted in 5,500 publications per month or 250 per workday over the two year period form May 2003 to March 2005. In its first two months of operation, the compliance rate was 340 submissions or 3 percent of the expected total. Available from: http://www.earlham.edu/~peters/fos/newsletter/08-02-05.htm.

months via NIH's PubMed Central online digital archive."[66] As of this writing, the bill is still pending. Updates about the proposed legislation will appear in ARL's *SPARC Open Access News*.[67] As of December 31, 2005, NIH had received 2,830 articles under NIHPA and 745 were available in PMC. By mid-February 2006, PMC held more than 1,600 NIHPA articles. According to PMC staff, the lag between submission and availability of these articles in PMC stems from two factors: (1) internal processing time, which is typically a few weeks, and (2) an author may delay release of an article in PMC for up to 12 months after publication.[68]

Meanwhile since July 2005, the Research Councils UK (RCUK) has promulgated an even more far-reaching draft policy that would make all government-funded research in the UK freely available to the public. While it has yet to be adopted, the biomedical community is already leading the way. In June 2004, the NLM announced a cooperative project with The Wellcome Trust, the UK's largest non-governmental funding source for biomedical research, and JISC (Joint Information Systems Committee) to digitize, and make freely available to the public, the complete backfiles of a number of historically significant research journals. Effective October 1, 2005 Wellcome Trust began to require public deposit of electronic copies of any research papers supported wholly or in part by its funding, within six months of publication.[69] In response, "Oxford University Press, Blackwell and Springer changed their copyright agreements with authors to allow immediate self-archiving of Wellcome-funded research."[70] PubMed Central (USPMC) serves as the central repository while a UK PubMed Central (UKPMC) is under development. The UKPMC site, which will serve as a mirror to USPMC while also accepting UK submissions, is expected to launch in early 2007 with more than 500,000 research articles. The UKPMC represents an alliance among six biomedical research and funding agencies, led by The Wellcome Trust.

PMC has eliminated the "SmartSearch" label discussed in the 2003 DLF survey; however, the underlying technology is still used. There are numerous improvements to PMC's search interface and functionality. PMC serves as one of many sources of full-text

---

[66] ARL, Federal Relations E-News, Winter 2006. Available from: http://www.arl.org/info/frn/frnmon.html.

[67] In *SPARC Open Access Newsletter*, issue 93 (January 2, 2006), Peter Suber provides background information about the U.S. CURES Act, including ways in which it goes beyond the current NIH public access policy. Available from: http://www.earlham.edu/~peters/fos/newsletter/01-02-06.htm.

[68] NIHPA data for December 2005 and February 2006 was provided in email communication with Ed Squeira on February 15, 2006.

[69] Refer to The Wellcome Trust's Web page on "Open and Unrestricted Access to the Outputs of Public Research" available from: http://www.wellcome.ac.uk/node3302.html.

[70] Article by Kate Worlock as cited by Peter Suber in *Open Access News* on December 23, 2005. Available from: http://www.earlham.edu/~peters/fos/2005_12_18_fosblogarchive.html

articles linked to PubMed and MEDLINE citations and the Entrez retrieval system supports access to online books, sequence databases, a taxonomy database and other resources. Users can search the full-text of all SGML or XML-based content deposited in PMC and there are various linking options across articles, issues and journals to commentaries, cited in, referenced articles and corrections. The PMC "Utilities" tab includes an "Open Access List" of journal titles included in PMC with fully or partially open content.[71]

Author manuscripts resulting from NIH's public access policy have a distinctive page banner and watermark with a left-margin stripe running the length of the record (Figure 20).

Figure 20: Example of a NIH Public Access Author Manuscript from PubMed Central (2/14/06)
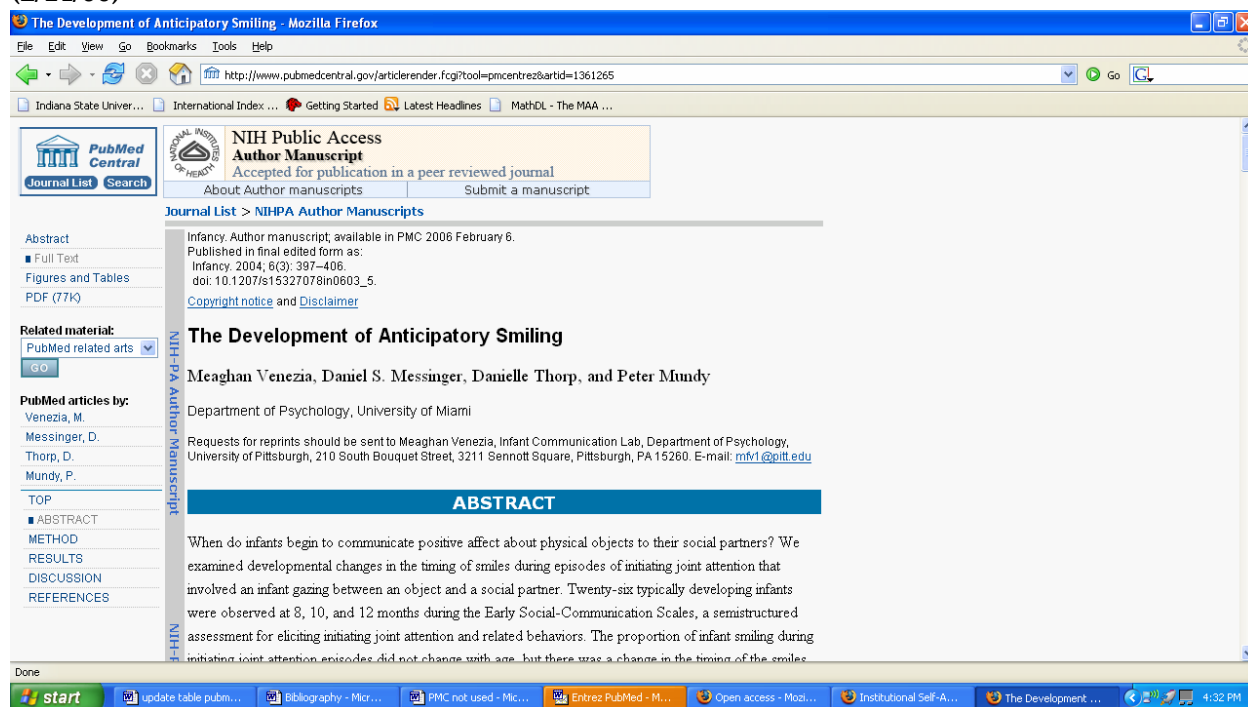
In early February 2006, NLM announced that it created a new status tag to PubMed citations, signaling author manuscripts for published articles added to PubMed Central due to the public access policy. According to the press release, the new status tag, [PubMed - author manuscript in PMC], "appears on PubMed citations for articles that would not normally be cited in PubMed because they are from journals that are a) not indexed for MEDLINE or b) do not participate in PMC. This small number of citations can be retrieved using the search: pubstatusnihms. As these citations are processed,"

---

[71] For these and other enhancements refer to Brooke Dine's PowerPoint presentation to the Medical Library Association on "PubMed Central," May 2004. Available from http://www.nlm.nih.gov/pubs/techbull/ja04/theater_ppt/pmc.ppt.

PubMed continues, "the status tag will change as appropriate, with a final designation of [PubMed]. To retrieve all citations in PubMed for which author manuscripts are available in PMC, use the search: author manuscript [sb]." [72] As of mid-February 2006 the PubMed search query "pubstatusnihms" retrieves 66 articles, whereas the later search query, "author manuscript [sb]" yields 1,655 results.

Figure 21: Record for the above article as it appears in PubMed (2/14/06)
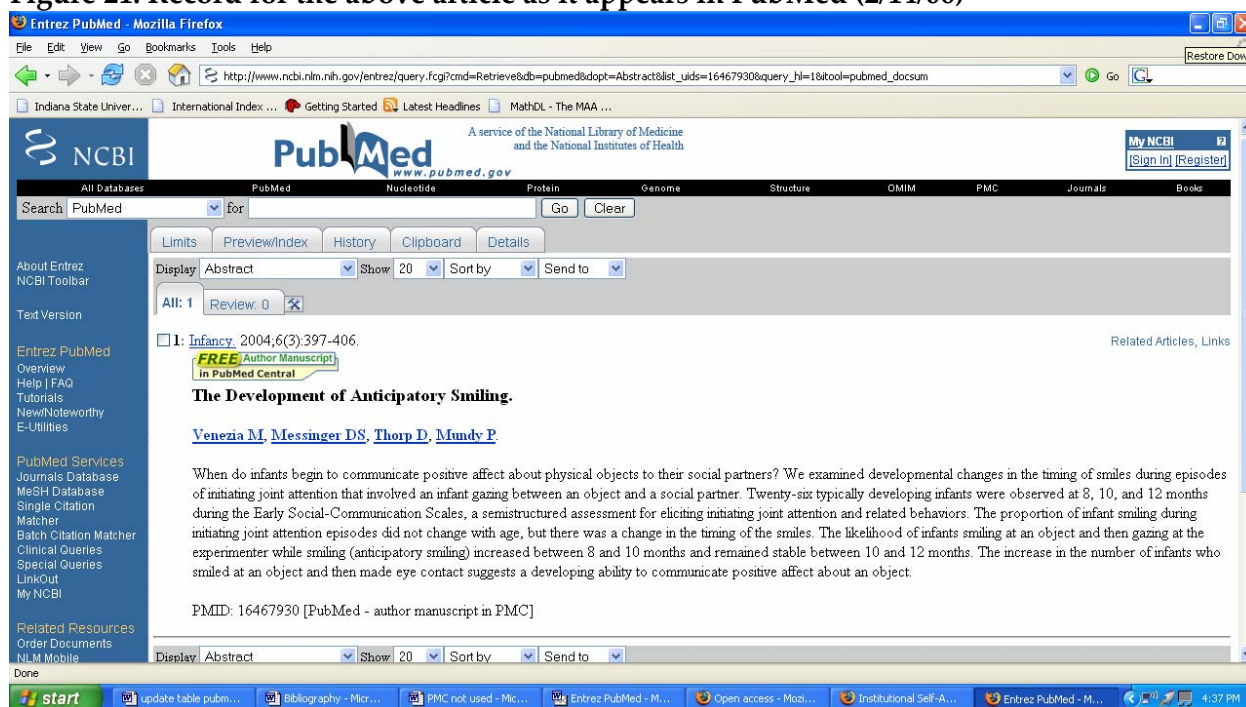
PubMed Central's phenomenal article retrieval statistics provide persuasive evidence that it attracts a wide spectrum of users. In an editorial discussing the impact of the *Journal of the Medical Library Association's (JMLA)* participation in PMC, T. Scott Plutchak revels in the increased exposure open access brings to the journal—an estimated 20,000 to 30,000 unique readers monthly or about four to six times the core audience of 4,500 MLA members (Plutchak 2005). However, when evaluating its potential impact on MLA membership and *JMLA* revenues, Plutchak is more tentative, stating that the "jury is still out," and that "it is too early to label the experiment [open access] an unqualified success." So far, the impressive usage statistics have persuaded him that open access is worth the risk. Only time (and revenues) will tell if the MLA will continue to support public access on a permanent basis.

---

[72] NLM *Technical Bulletin*, 348 (January-February 2006), posted on January 27, 2006. Available from: http://www.nlm.nih.gov/pubs/techbull/jf06/jf06_technote.html#note . Explanatory annotations about all of PubMed's citation status tags are available from: http://www.nlm.nih.gov/bsd/pm_cit_status.html. For background information refer to Sequeira (2005).

4.2.4 CDS: CERN Document Server

Update Table 08: CERN Document Server based on DLF Survey responses, Fall 2005

	CERN Document Server cds.cern.ch/
ORGANIZATIONAL MODEL	International organization
SUBJECT	Science
FUNCTION	High Energy Physics and related areas long-term archive and search engine
PRIMARY AUDIENCE	Research Community
STATUS	Established
SIZE	800,000 bibliographic records, including 360,000 full-text documents. 1,200 new documents added per week.
USE	Per day: 7,000 searches. Per month: 20,000 unique users.
ACCOMPLISHMENTS	1. Internationalization: 14 languages translated. 2. Word-frequency ranking, impact factor ranking, citation ranking, find similar records functionality. 3. OAI compliancy for both harvesting and providing metadata.
CHALLENGES	1. Impact measurement, combining the various ranking weights. 2. Collaborative tools to share baskets, alerts, annotations, comments, reviews, etc. 3. Extending CDSware technology to support up to 20 million records.
TOOLS OR RESOURCES NEEDED	Sharing impact values with repositories serving same documents (especially full-text download counts).
GOALS OF NEXT GENERATION RESOURCE	Complete digital library system, with Google-like features. OAI compatibility GPL distribution.

Founded in 1954, CERN, the European Organization for Nuclear Research with 20 European member states, constitutes the largest particle physics' laboratory in the world. For more than fifty years, CERN has been an international proponent of "publishing or making generally available" the research results of its experimental and theoretical work, as originally mandated by the CERN Convention (Pepe et al. 2006). Since its inception the CERN Library has operated a document archive and free preprint distribution service. Over the past twelve years, CERN Library services have evolved on the Web as an institutional repository, starting with dissemination of preprints, then extending access to periodicals, books and other library-related materials, and today, integrating all types of multimedia materials including photos, posters, lectures, and videos into the CERN Document Server (CDS). In addition to providing access to CERN documents, CDS harvests metadata from related subject repositories, including arXiv. (In fact, the majority of CDS full-text documents come from external sources. As of mid-February 2006, UIUC Grainger Engineering Library harvested an estimated 70,000

CERNmetadata records, but OAIster only harvests around 38,500 full-content items directly from CERN.)[73]

Besides hosting documents in the field of high-energy physics (HEP), CDS provides a growing suite of tools and services to facilitate sophisticated searching, collaborative and social networking, and citation and usage metrics (Pepe et al. 2006). The search interface offers the ability to limit queries by field or collection as well as manipulate search results through options in sorting, display and output. Marked results can be saved and stored if users register and log-in. CDS expects to "adopt a comprehensive system of commenting, reviewing and messaging that will allow users and groups to discuss content and share knowledge privately and publicly" (Pepe et al. 2006, 3). CDS also generates a citation index through the extraction of references from full-text documents and uses it to rank documents according to the number of times it is cited by or co-cited with other papers. Finally, the CDS system ranks documents based on the number of downloads and offers users links to "find similar" documents from each result.

Presently, CERN estimates that only 30 percent of its scientists' current article production is not available as open access on CDS; moreover, the library plans to fill this gap.[74] CERN has achieved this impressive record by steady implementation of practices and policies in support of open access, including adoption of the OAI protocol in 2002 and promulgation of an electronic publishing policy in 2003. The policy encourages:

- submission of all CERN scientific documents to a relevant e-archive;
- extension of electronic publishing across all forms of scholarly communication (e.g., conference proceedings);
- "publishing in low-cost, easily accessible electronic journals," taking into account "the publication costs and the subscription policy of the journal"; and,
- equal attribution of relevance to referred articles in electronic (as compared to traditional) journals when selecting candidates for positions at CERN. [75]

In 2004, CERN signed the Berlin Declaration making official its commitment to open access principles. By 2005, the field of particle physics could claim nearly 100 percent open access to research results through the combined initiatives of arXiv, the SPIRES HEP database (sponsored by the Stanford Linear Accelerator Center), and CDS. Despite

[73] See Yeomans (2006) for details about their efforts to obtain CERN-authored documents from varied sources.

[74] CERN Action on Open Access identifies milestones and links to major policy documents. Available from http://open-access.web.cern.ch/Open-Access/pp.html.

[75] "Continuing CERN action on Open Access," issued by the Scientific Information Policy Board, summarizes CERN open access policy development up to March 2005. Available from http://www.ecs.soton.ac.uk/~harnad/Temp/cern.pdf . An Electronic Publishing Policy for CERN (November 2003) is available from http://library.cern.ch/cern_publications/SIPBPubPol.17.11.03.htm .

the success of near global availability of pre- and post-prints in electronic archives, CERN officials observed that it had not engendered widespread adoption of new publishing models or altered criteria for academic advancement. The prospect of CERN's new flagship particle accelerator (Large Hadron Collider—LHC) launching in 2007, prompted CERN to initiate a series of high level meetings, bringing together major physics publishers, research laboratories, learned societies, funding agencies and authors to discuss transition strategies towards publishing models that would support open access and lower cost journals for LHC research results.[76] Participants disclosed:

> The LHC collaborations feel positive about exploring new publishing models provided that features such as peer-review and long-term archiving are preserved. It is also of high importance that the funding agencies start to consider publication costs as being part of research budgets. In addition, it was stressed that open access publishing requires a range of actors, as has been the case under the current paradigm, in order to regulate the market and maintain a healthy competition among the publishers.[77]

As a result of this meeting, CERN formed a task force mandated to bring about action by 2007. In the press release, CERN's Director General Robert Aymar gave this endorsement:

> The next phase of LCH experiments at CERN can be a catalyst for a rapid change in the particle physics communication system. CERN's articles are already freely available through its own web site but this is only a partial solution. We wish for the publishing and archiving systems to converge for a more efficient solution which will benefit the global particle physics community. [78]

CERN's *High Energy Physics Libraries Webzine*, freely accessible from the Library's Web site, features articles about recent developments at CDS and in the field in general. For example, an in-depth article about applying usage statistics to CERN's e-journal collection appeared in August 2005 (Dominguez) and two articles from March 2006 examine CERN's continuing participation in Open Access (Gentil-Beccot 2006) and investigate the growth in its metadata and full-text eprint coverage (Yeomans 2006).

---

[76] The first, "Open Meeting on the Changing Publishing Model," was held at CERN on September 16, 2005. Video streaming and minutes are available from http://open-access.web.cern.ch/Open-Access/20050916.html. The second meeting is referenced in the next footnote.

[77] "Colloquium on Open Access Publishing in Particle Physics" held at CERN on December 7-8, 2005. Summary is available from http://indico.cern.ch/conferenceDisplay.py?confId=482. Minutes of the meeting, including presentations, are available from http://indico.cern.ch/getFile.py/access?resId=0&materialId=minutes&confId=482

[78] European Organization for Nuclear Research. Press Release, December 14, 2005, "A Step Forward for Open Access Publishing." Available from http://press.web.cern.ch/press/PressReleases/Releases2005/PR18.05E.html

From 2001 to 2005, the CDS Library at CERN offered high-profile annual workshops on the Open Archives Initiative. Beginning in 2005, the CERN workshops are held every second year in alternation with the Nordic Conference on Scholarly Communication (2006). Presentations (and Web casts) from these conferences are available from their respective Web sites.[79]

4.2.5 OLAC: Open Language Archives Community

Update Table 09: OLAC based on DLF Survey responses, Fall 2005

	Open Language Archives Community (OLAC) http://www.language-archives.org/
ORGANIZATIONAL MODEL	International partnership of institutions and individuals
SUBJECT	Language resources
FUNCTION	Network of language archives conforming with the Open Archives Initiative; Virtual library
PRIMARY AUDIENCE	Academic Community
STATUS	Established
SIZE	28,000 records (41% increase) from 34 archives (36% increase)
USE	2005: 824,676 queries, an average of 2259 per day or an average 68273 per month.
ACCOMPLISHMENTS	1. Google-style search interface. 2. Report Card system for metadata quality. 3. Continued steady growth in participation.
CHALLENGES	1. Sponsorship for maintaining core services. 2. Lack of a good metadata editor.
TOOLS OR RESOURCES NEEDED	1. Publicity, profile. 2. Guidance on long-term funding sources other than research agencies. 3. More cool services based on OAI content.
GOALS OF NEXT GENERATION RESOURCE	Fully operational now but need to maintain service and continue to support wider adoption, metadata cleansing, etc.

OLAC continues to fulfill its twin stated objectives of developing: (1) consensus on best current practice for the digital archiving of language resources and (2) a network of interoperating repositories and services for housing and accessing such resources. It comprises an estimated 28,000 records aggregated from 34 participating archives. OLAC aims to provide linguists with the data, tools, and advice relevant to the study of human languages, documented in digital and non-digital form from published or restricted sources.

---

[79] See "Innovations in Scholarly Communication" (OAI4) held at CERN in October 2005, available from: http://indico.cern.ch/conferenceDisplay.py?confId=0514 . Information about the Nordic conferences held at Lund University are available from: http://www.lub.lu.se/ncsc2006/.

OLAC's founders continue to recruit new content by offering tutorials, making conference presentations, and participating in the interdisciplinary research community.[80] OLAC actively promotes the E-MELD (Electronic Metastructure for Endangered Languages Data) Project, funded for five years through 2006 by the National Science Foundation. Among other initiatives, E-MELD has created the successful online "School of Best Practices in Digital Language Documentation."[81] In summer 2006, E-MELD will host a Digital Tools Summit in Linguistics, held in conjunction with the Linguistic Society of America (LSA) conference. The summit will address the cyberinfrastructure needs of linguistics and extend the work of the "E-MELD Toolroom."[82] According to the summit's organizers:

> Linguistics is at a critical moment, as the need for more accurate, re-useable and typologically diverse data, together with the increasing urgency of worldwide language documentation, converge to drive the development of digital tools and cyberinfrastructure. Access to language corpora has become indispensable for a wide range of linguistic inquiry, including basic research (in e.g. phonetics and phonology, syntax, semantics and morphology, psycholinguistics and language documentation) and applied research (in e.g. speech engineering, sociolinguistic modeling, language revitalization and pedagogical materials development). This use of large and small corpora to conduct research on both well-documented and poorly-documented language varieties has resulted in the emergence of a new interdisciplinary confluence of computational linguistics, language documentation, and linguistic theory.

> Linguists and language developers have particular challenges in developing high-quality, exchangeable, re-useable corpora: standards and tools for encoding and rendering, annotation, querying, archiving, and generating presentation formats are all in their infancy. Linguists' materials often include multiple modes of media, multiple languages, and multiple levels of analysis.[83]

Four new archives joined OLAC in 2005. While the number of archives represented in OLAC has increased, half of the content is derived from SIL International (formerly Summer Institute of Linguistics), specifically from SIL: Language and Culture Archives (metadata records extracted from the bibliography of 20,000 citations spanning 70 years of SIL International's language research in over 2,000 languages) and from Ethnologue: Languages of the World (metadata records for each of the 7,000-plus modern languages

---

[80] See for example, Gary Simons' tutorial, The Open Language Archives Community: Building a Worldwide Library of Digital Language Resources, (January 2006) available from http://www.language-archives.org/events/olac05/.

[81] E-MELD School of Best Practices in Digital Language Documentation is available from http://emeld.org/school/index.html.

[82] EMELD Toolroom is available from http://emeld.org/school/toolroom/index.html.

[83] Information about the summit is available from http://www.ku.edu/pri/DTSL/.

in the world—both living and recently extinct—identified in the Web edition of this reference work). Two other sources—Digital Archive of Research Papers in Computational Linguistics and PARADISEC (Pacific and Regional Archive for Digital Sources in Endangered Cultures), each contributing more than 3,000 records—make up an additional 25 percent of the content. A number of archives have remained static in size since 2003; 15 archives contribute fewer than 25 records.[84]

OLAC provides a useful synopsis of search queries for 2005:

> In 2005, the OLAC Search Engine handled 824,676 queries, an average of 2259 per day or an average 68273 per month. The most popular languages searched for in 2005 were Dutch, English, Quechua, Arabic, Greek, German, Chinese, and Malay. Only 35 percent of queries specified a particular archive, the majority were generic searches across all archives. The most commonly searched repository was SIL-LCA, followed by PARADISEC and SCOIL.

An early adopter of OAI, OLAC operates a registry service with guidelines about how to become a data provider (including static repository implementations); conformance testing and validation of new archives is integral to the registration process. OLAC supports a unique metadata standard, based on all 15 elements of Dublin Core, supplemented by metadata extensions with controlled vocabularies specific to the community, including Language Identification, Linguistic Data Type, Linguistic Field, Participant Role, and Discourse Type.[85]

Since 2003, OLAC developers have introduced an innovative "metadata report card" system to assess the semantic and syntactic quality (as opposed to the structural composition) of the metadata submitted by each archive (Hughes 2004).[86] According to the composite Archive Report Card, OLAC receives a score of 6.77 out of 10 points for metadata quality, with an average of 8.77 elements per records.[87] Every individual archive is also given a score according to its conformance to OLAC's metadata best practice guidelines. For example, the 3,018 records in PARADISEC have an average of 10.71 elements per record and receive an average score of 7.99/10 for metadata quality taking into consideration the usage of elements and codes in the record.
Figure 22: Screenshot of PARADISEC's Archive Report Card in OLAC

---

[84] Participating Archives are listed at http://www.language-archives.org/archives.php4.
[85] OLAC Metadata explanation is available from http://www.language-archives.org/OLAC/metadata.html.
[86] OLAC metadata report system is available from http://www.language-archives.org/tools/reports/ExplainReport.html.
[87] The composite OLAC Archive Report is available from http://www.language-archives.org/tools/reports/archiveReportCard.php?archive=all.

Source: http://www.language-archives.org/tools/reports/archiveReportCard.php?archive=18
(April 18, 2006)

To inform their "efforts to create good controlled vocabularies," OLAC has also initiated a survey of OLAC metadata implementations that allows "users to see how any attribute or field of OLAC metadata has been used by OLAC archives." [88]

---

[88] OLAC Metadata Survey is available from http://www.language-archives.org/tools/survey.php4 .

Figure 23: Screenshot of OLAC's Metadata Survey Tool
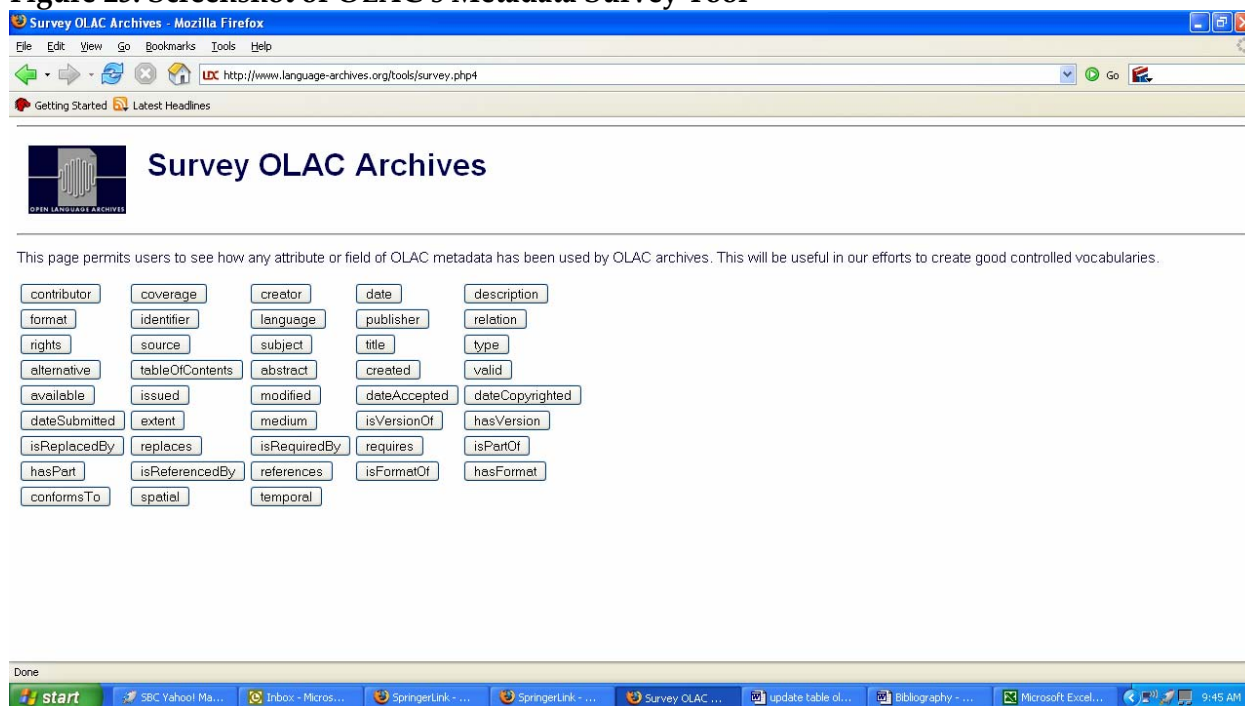
Source: http://www.language-archives.org/tools/survey.php4, (April 8, 2006)

Clicking on any element in the survey brings up details about the frequency, language, type, code, and content with which it is used. For example, as indicated below in the top results for "contributor" in OLAC, Arthur Capell is identified as a researcher 859 times and as an author 401 times in OLAC.

Element: contributor

freq	lang	type	code	Content
859		role	researcher	Capell, Arthur
851		role	depositor	Newton, Peter
401		role	Author	Capell, Arthur
314		role	recorder	Durie, Mark
225		role	recorder	Dutton, Tom
183		role	recorder	Voorhoeve, C.L.
162		role	photographer	Thieberger, Nicholas

Among OLAC's most significant accomplishments since 2003 is its implementation of a Google-style search interface (http://www.language-archives.org/tools/search/). Searches can be conducted across the entire aggregation or limited to specific archives.

Features of the search engine include a variety of string matching algorithms; a thesaurus of alternate language names; language code searching; keyword-in-context display in search results; search for similarly spelled words; search for similar items; support for standard string search operators and domain-specific inline syntax; and automatically derived search links for other web search engines. A notable contribution of this research is the inclusion in the search engine results of a metadata quality-centric sorting algorithm (Hughes and Kamat 2005).

Figure 24: Screenshot of Results for <eskimo> in OLAC
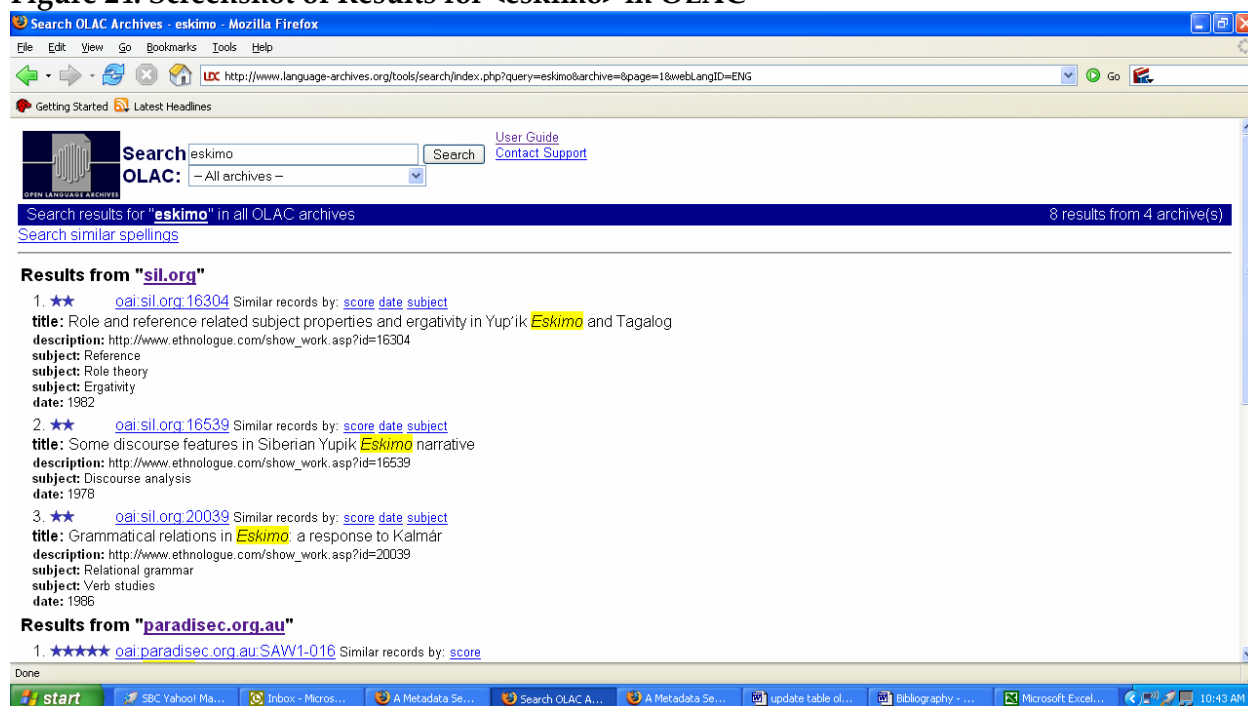
Source : http://www.language-archives.org/, (April 8, 2006)

OLAC has also implemented a customized version of an OAI DP9 gateway for Web crawlers, facilitating the indexing of its constituent archives' Web pages by generic Internet search engines.

In addition, as described in the original DLF survey, OLAC is searchable via The Linguist List Web site (in basic and advanced search modes).[89] Full documentation about OLAC is available from its Web site, http://www.language-archives.org/documents.html.

---

[89] The Linguist List is available from http://linguistlist.org/olac/.

4.2.6 Electronic Theses and Dissertations (ETDs)

ETDs continue to figure heavily in the content of e-print repositories and frequently serve as a core component of university IR deployment strategies. The Networked Digital Library of Theses and Dissertations (NDLTD), celebrating its tenth anniversary in 2006, is an international federation that aims to improve graduate education by developing accessible digital libraries of theses and dissertations. NDLTD charges annual dues for membership based on institutional configuration (single degree-granting versus multi-campus systems and consortia) and country of origin. The 2003 United Nations Human Development Report is used to group countries into three categories. As a result, membership fees vary widely, ranging from $100 per year for a single institution from a category II or III country to $75,300 for a consortium with 500 or more members from a category I country.[90]

The NDLTD Membership Directory is accessible at its Web site and can be sorted by country, name of institution, last update, and join date.[91] Every institution is linked to a template that provides more details about its deployment of ETDs, including the number collected, their formats and languages, search and retrieval information, catalog access (OPACs), and organizational contact information. Unfortunately, many institutions have incomplete and out-of-date information. This skews the composite statistics, which would otherwise be quite valuable. Throughout the six-month duration of writing this report, NDLTD's membership site was in transition. There are 231 NDLTD members, constituting 201 member universities (including 7 consortia) and 30 institutions.[92]

UNESCO's "Guide to Electronic Theses and Dissertations" (2001) and an online ETD tutorial developed by Ohio State University in cooperation with Adobe Acrobat, Inc. and NDLTD (Gray et al. 2005) are available from NDLTD's Web site along with links to NDLTD's annual conference information.[93] The site's wiki, launched in October 2005, contains basic documentation, but has not been fully developed as of mid-May 2006.

---

[90] The structure of dues is outlined at NDLTD's Web site: http://www.ndltd.org/join/dues.

[91] NDLTD's Membership Directory available from http://www.ndltd.org/membership/dir.html is not actively maintained.

[92] Information available from http://tennessee.cc.vt.edu/~lming/cgi-bin/ODL/nm-ui/members/. Links to university sites not functioning at the time this site was consulted. Break-out figures about ARL members, etc. are also out of date.

[93] The UNESCO Guide (420 pages in length) is intended to be a "living document" with regular updates; however, it does not appear to have been refreshed recently. It is available from http://www.ndltd.org/etdguide/ETDGuide.pdf and http://www.etdguide.org/. The ETD Tutorial is available from http://etd.vt.edu/etdtutorials/ and http://www.adobe.com/education/etd/tutorials.html.

Using NDLTD's membership information and OpenDOAR (described in section 4.1.5) as points of comparison, there are more than 200 institutions in over 40 different countries actively collecting electronic theses and dissertations. Among the 380 repositories registered at OpenDOAR as of mid-May 2006, 227 contain dissertations (and 78 have undergraduate theses). In terms of content types, dissertations are second only to "articles" (with 241 instantiations) in OpenDOAR. Of course these figures give an incomplete picture since not all institutions with ETD deployments are NDLTD members (OhioLink for example dropped its consortial membership in NDLTD, leaving it up to individual institutions to join if they wish) nor is OpenDOAR comprehensive. There are a growing number of ETD aggregations organized at the state, national, and trans-national level. Many such efforts build on long-standing traditions of coordinated bibliographic control of citations and abstracts of theses and dissertations in print form. Notable examples are noted below.

- OhioLink has a growing online catalog of electronic theses and dissertations from member institutions that includes full-text (circa 7,300) when available. Accessible from: http://www.ohiolink.edu/etd/.
- The Theses Canada Portal, hosted by the Library and Archives of Canada, (launched in January 2004), contains nearly 46,000 full-text ETDs as of mid-May 2006.[94] Accessible from http://www.collectionscanada.ca/thesescanada/index-e.html.
- In Africa, the Association of African Universities maintains the Database of African Theses and Dissertations (DATAD) which includes a growing number of full-text ETDs. Accessible from http://www.aau.org/datad/database/.
- In Brazil, the IBICT (Instituto Brasileiro de Informação em Ciência e Tecnologia), funded by the Ministry of Science and Technology, coordinates the library of Brazilian digital theses and dissertations, BTDT (Biblioteca Digital de Teses e Dissertações). Accessible from http://bdtd.ibict.br/.
- The original gateway to Australian digital theses, (initiated by seven universities in 1998), expanded in scope and re-launched in January 2006 to become Australasian Digital Theses, embracing institutions in Australia and New Zealand.[95] To help bring more repositories online, ADT partnered with ProQuest in 2005 to test its Digital Commons Deposit and Repository software (Kennan et al. 2005). This initiative not only brought new content into the aggregation but also increased its user base. (In early 2006, ADT contained about 3,500 ETDs). Nevertheless, a recent study of the impact of mandatory ETD policies in Australia concludes that universities "seem to be wasting their money if they maintain a voluntary deposit policy" and further finds that mandatory policies based on date of submission achieve 80 percent compliance rates five or six years

---

[94] The ETD search page is available from http://amicus.collectionscanada.ca/s4-bin/Main/BasicSearch?coll=18&l=0&v=1.

[95] Information about ADT's redeployments is available from http://adt.caul.edu.au/newsprojects/adtariic/.

faster than policies dated from enrollment (Sale 2006).   For ADT to succeed, Sale
advises that it must advocate strongly for mandatory thesis submission policies.
He further suggests that if the Australian Government amended its guidelines
for Australian Postgraduate Awards (APAs) by requiring graduates to deposit
both a paper and an electronic copy of his or her thesis with the university, the
ETD deposit rate would increase dramatically as universities would likely extend
the requirement to all graduates. (Accessible from http://adt.caul.edu.au/.)

On the European front, JISC (UK) and SURF (Netherlands) convened a workshop in
January 2006 to discuss ETD trends and issues. The pre-conference survey responses
from 11 countries reveal how different cultural, educational, governmental, and legal
frameworks impact the deployment of systematic ETD programs in various European
countries.[96] Many countries have active centrally-managed ETD programs underway.
Especially noteworthy are:

- The Scandinavian DiVA (Digitala Vetenskapliga Arkivet) portal; accessible from
  http://uppsok.libris.kb.se/sru/uppsok.
- The Netherlands' "Promise of Science" initiative to make 10,000 e-these available
  by the end of 2006 as part of its national search and discovery service, DAREnet.
  Accessible from (http://www.darenet.nl/nl/page/language.view/promise.page.
- The EThOS Project in the UK, co-supported by the British Library and the
  Consortium of University Research Libraries (CURL), which aims to develop a
  prototype e-theses service in the framework of a national infrastructure.
  Accessible from http://www.ethos.ac.uk/;
- DissOnline, Digital Dissertations on the Internet, coordinated by the German
  National Library, where a portal is under development that will enable the
  integration of domain-specific subsets into the German interdisciplinary Internet
  for scholarly information, Vascoda (http://www.vascoda.de/) or other services.
  Accessible from http://www.dissonline.de/.

Jacobs (2006a) provides a useful summary of the workshop's findings relative to
national trends and elaborates on common thematic issues, including site
interoperability, enrichment (links to data/multimedia, and preservation), and
management. He notes that a fundamental, and unresolved, issue revolves around
whether or not ETDs warrant a separate pan-European gateway or if they should be
treated as a manifestation of research output alongside many others, and integrated into
a generic European repository infrastructure. Two prototypes under development,
DART-Europe and DRIVER, represent these different conceptual approaches. DART-
Europe is briefly described below and DRIVER is discussed more fully in section 2.1.
(See Figure 03).

---

[96] All presentations are available from http://www.surf.nl/bijeenkomsten/index6.php?oid=203.
See in particular Neil Jacobs' "Overview of all countries,"
http://www.surf.nl/download/Overview%20of%20all%20countries.pdf.

- DART-Europe (Digital Access to Research Theses-Europe)
  http://www.dart-europe.org/
  The University College London and Dartington College of Arts in partnership with ProQuest are exploring the creation of a pan-European portal for the "deposit, discovery, use and long-term care of research theses." The DART-Europe gateway would enable "free-at-the-point-of-use access to the full text of electronic research theses," leveraging ETD efforts at the institutional and national level across Europe. DART-Europe intends to move beyond "traditional, text-based material" to embrace "disciplines and institutions that are already widening the definition of research by redefining the formats of theses." Project plans and conference papers are available at the Web site.

ETD Software Developments

There is no comprehensive source to compare what software systems are used by ETD services worldwide. At present, the best source is ROAR (described in section 4.1.4). Despite its incomplete representation, ROAR offers easy identification of OA repositories with e-theses content through drop-down menus with filters by software system and country. Among the 68 e-theses archives from 19 countries (there are no entries from the U.S. in this category) listed in ROAR in mid-May 2006, Virginia Tech's ETD-db software has the greatest use (23), followed by GNU EPrints (13), DSpace (7), and the Danish DoKS system (3). Two other systems have one instantiation: the Swiss CDSware and the German MyCoRE. Twenty other deployments use various "other" software programs. The base URLs for these systems as well as Fedora are provided below along with links to relevant background information.

- ETD-db: http://scholar.lib.vt.edu/ETD-db/#about
- GNU EPrints: http://www.eprints.org/software/
  - o How to Create a Theses Repository:
    http://www.eprints.org/software/howto/theses/
- DSpace: http://www.dspace.org/
  - o Theses Alive Plug-In for Institutional Repositories (TAPIR) developed at Edinburgh University Library (Jones 2004):
    http://www.ariadne.ac.uk/issue41/jones/
  - o TAPIR SourceForge page: http://sourceforge.net/projects/tapir-eul/
- DoKS: http://www.doks.dk/
- CDSware: http://cdsware.cern.ch/
- MyCoRE: http://www.mycore.de/engl/
- Fedora™: http://www.fedora.info/
  - o VALET from VTLS: http://www.vtls.com/Products/valet-for-ETDs.shtml
    VALET for ETDs is a customizable, web-based interface that allows remote users to submit Electronic Theses & Dissertations into a FEDORA™ digital object

repository. . . VALET for ETDs is offered as a free, open-source solution for web self-submission of ETDs. This solution builds upon our collaborative experience with the NDLTD Project at Virginia Tech, the ADT Program and the ARROW Project in Australia to build a 'best of breed' solution for web submission of ETDs.

ETD Union Catalogs and Search Engines

OCLC runs the NDLTD Union Catalog using OAI with a SRU service on top. As of mid-May 2006, it harvests 242,458 ETD records from 59 entities, wherein the OCLC ETD set is the largest, with 69,564 records, followed by the Library and Archives Canada ETD repository with 45,795 records.[97] The catalog has an undetermined number of duplicate records. OCLC harvests all relevant ETD records including bibliographic metadata only as well as records with associated abstracts or full text, when available. It is estimated that about 70 percent of the records in the NDLTD Union Catalog represent unique full-text ETDs.

A variety of services have built search engines based on the NDLTD Union Catalog's harvested records. Among the six options to "browse or search ETDs" linked from NDLTD's Web site (http://www.ndltd.org/browse.en.html), several have restricted access, are out-of-date, or only represent a subset of available ETDs. Among the useful links for digital library developers is the SRU search by OCLC . This Web service machine interface performs remote searches through OCLC's central NDLTD metadata collection. External services and portals may directly connect to this service for seamless integration of ETD searching into any other system (http://alcme.oclc.org/ndltd/SearchbySru.html).[98]

The VTLS deployment (http://zippo.vtls.com/cgi-bin/ndltd/chameleon) updated in March 2006 after a long period of dormancy contains 160,392 records, virtually all of which have associated URLs to the full text. The user interface can be switched to appear in ten languages other than English, including a number of Slavic languages and other non-Roman scripts, such as Arabic and Korean. The ETD content itself appears in more than 25 languages; however, the vast majority is in English (135,000).[99] Users can save search results for printing or e-mailing, and it is possible to review "search history."

---

[97] OCLC maintains its harvesting statistics for the NDLTD Union Catalog at http://alcme.oclc.org/ndltd/servlet/OAIHandler?verb=ListSets.

[98] OCLC also runs an OAI viewer service against the NDLTD Union Catalog that enables filtering of sets and formats: http://errol.oclc.org/ndltd.oclc.org.html.

[99] The VTLS site contains no statistical information. Figures were provided in email correspondence with Vinod Chachra on March 6, 2006.

4.2.6.1 Scirus ETD Search Engine

Update Table 10: Scirus ETD Search Engine based on DLF Survey responses, Fall 2005

	Scirus ETD Search http://www.ndltd.org/serviceproviders/scirus
ORGANIZATIONAL MODEL	Partnership
SUBJECT	Cross-disciplinary
FUNCTION	Increase the visibility and accessibility of the content made available via NDLTD.
PRIMARY AUDIENCE	Research community
STATUS	Established
SIZE	220,000 ETDs
USE	Launched in October 2005
ACCOMPLISHMENTS	1. Increased visibility of NDLTD content. 2. Improved searching on NDLTD site. 3. Increase usage of NDLTD site.
CHALLENGES	1. Reliance on NDLTD Union Catalog for indexing and updates potential impediment.
TOOLS OR RESOURCES NEEDED	Indexing the NDLTD member sites individually rather than collectively via the Union Catalog.
GOALS OF NEXT GENERATION RESOURCE	No response.

The Scirus ETD search engine offers basic keyword or advanced searches that allow Boolean operators and fielded queries (among the 11 options are keyword, author, title, date, language, abstract, and rights). Searches can be limited to a specified range of publication dates, all subjects or any of 19 different subjects. Queries can be conducted within the ETD collection or the "Scholarly Web" at large.

Unlike its broader scientific search engine (Scirus described in section 4.5.1), Scirus ETD does not provide any data about its sources, size or scope, other than noting that the data is harvested from the NDLTD Union Archive hosted by OCLC. There is no "Help" page but a query for <ALL THE WORDS: global warming> retrieves 1,257 hits and is automatically rewritten as "global warming." Searches can be refined via a linked list to keywords found in the results. Results can be sorted by relevant or date (descending order only) and users can jump to the next page or in 20 page increments but not to the end of the results. There is no functionality to reorganize the chronological display or to save, store or email results.

Worthy of wider attention, is OhioLink's Worldwide ETD search service (http://search.ohiolink.edu/etd/world.cgi). The index is developed primarily from OAI harvesting of collections covered by the NDLTD Union Catalog. Records that appear only to have ETDs available for sale or accessible only on local campuses are removed.[100]

---

[100] Information provided by Thomas Dowling in email correspondence of February 24, 2006.

In addition, the service uses a Web crawler to retrieve a handful of sizeable ETD collections that do not have an OAI service, but which run on Virginia Tech's ETD-db software. As of mid-May 2006, the index contains almost 160,000 full-text, freely available ETDs. The user interface supports field-specific keyword searches. Searches can be limited to retrieve ETDs in "English only" or by level of degree (doctoral or masters). There are several options to sort results for display, but no post-processing features are available.

4.2.7 Grainger Engineering OAI Aggregation (UIUC)

Update Table 11: Grainger Engineering Library OAI Aggregation based on DLF Survey responses, Fall 2005

	Grainger Engineering Library OAI Aggregation http://g118.grainger.uiuc.edu/engroai/
ORGANIZATIONAL MODEL	Maintained solely by UIUC
SUBJECT	Science: engineering, computer science, physics
FUNCTION	OAI metadata harvesting aggregator
PRIMARY AUDIENCE	Research community
STATUS	Updating processing workflows
SIZE	672,000 records (52% increase) from 38 collections (tripled in number)
USE	Not tracking
ACCOMPLISHMENTS	1. Incorporation of additional OAI data services. 2. Improvements to workflow automation, including OAI harvester and indexer. 3. Lessons learned incorporated into other UIUC OAI projects
CHALLENGES	1. Scalability beyond one million records. 2. Human resources. 3. Data quality.
TOOLS OR RESOURCES NEEDED	1. More automated harvesting & aggregation tools.
GOALS OF NEXT GENERATION RESOURCE	Continue to grow to include all major physics, computer science, and engineering related OAI data sources

Grainger Engineering Library's OAI aggregation comprises a growing collection of e-prints, technical reports, theses and dissertations, and e-journals predominantly in the fields of engineering, computer science, and physics. The site is accessible from the Grainger Library's "Public Access Menu" (see "OAI Engineering Collection" under Technical Reports), but curiously missing as an option from its main "Resources" page. The OAI aggregation is a target in the Grainger federated search system called Grainger Search Aid, accessible from http://shiva.grainger.uiuc.edu/searchaid/searchaid.asp. Selecting the "Preprints and Open Reports" check box under Technical Reports and Preprint Servers in the left-hand frame will include this aggregation in the federated search.

As of May 2006, it covers 51 data providers and has more than one million items. It harvests records from many other services discussed in this report, including arXiv, CERN Document Server, DOAJ articles, OSTI's OAI repository, NSDL, Wolfram Functions, the Max Planck Institute, and UIUC's engineering document collection (16,300 items).

The utilitarian user interface and functionality have changed little since 2003 but it is now possible to display up to 500 short results per page as well as track queries during a session with the "search history" feature. There is no help page or advice about how to construct queries but the search syntax is returned with the results. A search for <carl lagoze> in the author/editor field returns ten results for "carl" near "lagoze"—all relevant and drawn from four different source archives (ECS e-prints, dLIST, Cogprints and arXiv).

In welcome contrast to many other services under review in this report, users can readily access data about the most recent OAI harvests via a link appearing at the bottom of the search page (http://g118.grainger.uiuc.edu/engroai/LastHarvest.asp). The intended frequency of harvests is not indicated but as of early May 2006, metadata from the majority of data providers has been re-harvested within the past three weeks. A handful of services, constituting more than 300,000 records, have not been re-harvested since January 2006 or earlier.[101] The total record count includes an undetermined number of duplicates.

4.2.8 PerX: Pilot Engineering Repository Xsearch

PerX, a cross-repository search tool focusing on engineering funded by JISC, is the result of a discipline-based landscape analysis and subject-specific inventory of relevant sources (http://www.engineering.ac.uk/). As discussed in section 2.1.2 of this report, the project's methodology, analytic framework, and deliverables are applicable to other disciplines. The pilot search service supports basic and advanced searches. In basic mode, the keyword query box is supplemented by a drop-down menu of options to limit the search by type of resource.

> Articles
> Theses & Dissertations
> Technical Reports

---

[101] All repositories are incrementally harvested twice a month starting every other day (to allow longer jobs to run to completion). In addition all repositories are scheduled to be fully harvested once every three months. Regular harvesting resumed on April 15, 2006 so the repositories that have not been re-harvested since January 2006 should be re-harvested by May 15, 2006. Official harvesting schedule provided in email correspondence with Thomas Habing, UIUC, on May 9, 2006.

Books
Learning & Teaching Resources
Key Web sites
Industry News
New Job Announcements
All

Advanced search mode supports Boolean operators and limiting to specific collections. A search in across all collections for <nanotechnology> returns 2,917 items, summarized by collection with an option to link the results. In this particular search, the most results (1,295) come from the COPAC union catalog, representing the holdings of 24 research university libraries in the UK and Ireland, plus the British Library, the National Library of Scotland and National Library of Wales (http://copac.ac.uk). Clicking on COPAC returns brief item records with the option to view the full record. When available, items are linked to full text. Users can control whether or not search terms are highlighted by turning the "highlight" function on or off. At this juncture there are no post-processing features other than the ability to edit searches, view the last result set or previous search queries.

PerX effectively demonstrates how to combine searches across different foundational resource collections (e.g., library catalogs, Web sites, learning object repositories) into a unified search interface.

4.2.9 CiteSeer

Update Table 12: CiteSeer based on DLF Survey responses, Fall 2005

	CiteSeer citeseer.ist.psu.edu/
ORGANIZATIONAL MODEL	Hosted by the PSU College of Information Sciences & Technology. Funded by NSF, NASA and Microsoft Research. Mirrors at U. of Zurich, MIT, & Nat'l U of Singapore. Linked to DBLP, ACM Digital Library, & SmealSearch.
SUBJECT	Computer & Information Science
FUNCTION	Search engine and digital library. Metadata resource. Open access to author-provided and self-archived documents.
PRIMARY AUDIENCE	Academic, Research and Educators
STATUS	Established but next generation under development.
SIZE	> 700,000 documents
USE	Per day: half a million hits with 20-50K documents downloaded; Per month: half million unique users.
ACCOMPLISHMENTS	1. Metadata extracted and available. 2. Mirrors established throughout the world. 3. CiteSeer model extended to SmealSearch, academic business. 4. Google Scholar instantiation of the CiteSeer model.

CHALLENGES	1. Scaling for millions of documents.
	2. New metadata indexed with new database.
	3. CiteSeer as a Web Service.
	4. Scalable modular architecture.
TOOLS OR RESOURCES NEEDED	1. More open source digital library and web resources.
	2. Funding to support continued development.
GOALS OF NEXT GENERATION RESOURCE	To build on the previous work of CiteSeer, expanding the service by increasing the breadth of the collection, and increasing and improving the site usability and services. To promote other such services.

Originally created at NEC Research Institute (now NEC Laboratories) by Steve Lawrence, Lee Giles and Kurt Bollacker, CiteSeer is hosted by Pennsylvania State University's College of Information Sciences and Technology with funding from NSF, NASA, and Microsoft Research. Comprising more than 700,000 records, CiteSeer is an autonomous citation index for computer and information science, created primarily through author and archive submissions, and Web crawling, using data mining and intelligent search functions.

The recipient of a $1.2 million NSF grant in mid-2005, Penn State and University of Kansas researchers will improve CiteSeer over the next four years.[102] According to Principal Investigator, Lee Giles, the next generation CiteSeer project will increase the breadth of the collection and enhance the site's usability and services. Giles outlines the following goals:

- To redesign the CiteSeer architecture for increased utility, reliability and services making it completely modular and open source.
- To expand the index to authors, affiliations, acknowledgements and others.
- To expand the breadth and depth of CiteSeer's collection.
- To have CiteSeer serve as a Web service for research use.
- To facilitate personalized CiteSeer search through the use of individual search histories combined with exploiting patterns of citations and searches within the community of users.
- To support collaborative CiteSeer usage and thereby to promote the formation and activity of research communities.
- To evaluate the impact of the new architecture, new content, and new services on the user community.
- To increase the reliability and sustainability of CiteSeer as a community resource.[103]

---

[102] Refer to the press release at Penn State Live, "Penn State IST researchers to enhance search engine" (August 29, 2005) available from http://live.psu.edu/story/13209.
[103] DLF Online Survey response, Fall 2005.

During the first five months of 2006, CiteSeer was unstable, preventing the author from being able to test its functionality effectively. However, CiteSeer has already expanded its search capability to include its new parsing service, which permits extraction of acknowledgments and header analysis. As a result, CiteSeer can now be searched by document (full-text source documents in PDF or PostScript formats), citation or acknowledgment.

A search conducted in early May 2006 of <open archives initiative> returns 64 articles, sorted in descending order by citedness. Lagoze and Van de Sompel's 2001 article, "The Open Archives Initiative: Building a Low-Barrier Interoperability Framework," heads the list with 18 citations. Clicking on the title launches a page with an abstract, offers links to the full-text document in various formats, and links to other citation indicators (citing and cited references), including a graph of citations to the article by year (only up-to-date through 2003). Similar articles based on the text and related articles based on references (co-citation) are generated. References within the original article are listed in citedness order and users can click on any title to identify other articles referencing this citation as well as *the context* in which the citations occur. Users can rate and comment on articles; they can also submit corrections.

In its present stage of development, CiteSeer is not without its glitches. Lagoze and Van de Sompel's article is noted with 18 citations on the opening results page, but with 19 on the detailed page. Moreover, when the author tried to retrieve these 18/19 citations, only seven were available, two of which point to the same article. HELP is available only from CiteSeer's companion search service for business, SMEALSearch, (http://smealsearch1.psu.edu/help/help.html). This page provides basic information about how to construct search queries (advanced and wildcard searches are not supported); describes the user interface; and answers frequently asked questions about algorithms, author contributions, document formats, legal issues and other matters.

In his highly favorable review of CiteSeer, Jacsó (2005a) concludes:

> What it lacks in user friendliness it makes up in smartness, especially in selecting high-quality sources, and in normalizing/standardizing the terribly inconsistent, incomplete and inaccurate citations prevalent in every scholarly field.

Effective February 2005, CiteSeer links to the ACM (Association for Computing Machinery) and DBLP servers.[104] Based at the University of Trier, the DBLP (Digital Library and Database Project) provides bibliographic information from major computer science journals and proceedings.[105] And more recently, Microsoft's new Windows Live

---

[104] See announcement available from http://citeseer.ist.psu.edu/announcements.html.

[105] Petricek and his colleagues investigated the differences and similarities between DBLP and CiteSeer, which rely on different methods of acquiring computer science literature. DBLP entries are entered manually whereas CiteSeer's are obtained by autonomous Web crawls and automatic

Academic search system, launched in April 2006, links to CiteSeer's content (http://academic.live.com/).

4.2.10   Citebase

Update Table 13: Citebase based on DLF Survey responses, Fall 2005

	Citebase http://www.citebase.org/
ORGANIZATIONAL MODEL	University of Southampton
SUBJECT	Science
FUNCTION	Online services, resources and tools to support self-archiving movement.
PRIMARY AUDIENCE	Research Community
STATUS	Experimental research service demonstration site
SIZE	370,000 documents (83% growth)
USE	Per day: 7,000 users
ACCOMPLISHMENTS	1. Metadata extracted and available. 2. 3 million linked references. 3. Easy to use interface with some novel features. 4. Linked from arXiv.
CHALLENGES	1. Scaling to other domains. 2. Scaling to usage and content. 3. Reduction of bugs and downtime. 4. Exit-strategy and sustainability.
TOOLS OR RESOURCES NEEDED	1. Structural improvement to code. 2. Publication and development of open source tools for citation linking. 3. Standardization of access to pen access full-text resources.
GOALS OF NEXT GENERATION RESOURCE	Cross-domain functionality. Transparency, user configuration, and author contribution.

Citebase, a prototype citation analysis service developed at the University of Southampton, is "an autonomous scientometric tool to explore and demonstrate the potential of OA material" (Hardy et al. 2005, 55). In his 2004 review of Citebase, Jacsó praised it highly asserting that Citebase:

---

processing of user submissions. Their research revealed that CiteSeer contains considerably fewer single author papers and is also biased against low-cited papers. Nevertheless, both databases presented similar citation distribution patterns. In comparing Computer Science citation patterns to Physics, the authors found that "highly cited papers in Computer Science receive a larger citation share than in Physics" (Petricek et al. 2005, 448).

> …shows the perfect model for the ultimate advantages of not only self-archiving scholarly documents but also of linking to full text – and offering citation/impact analysis on the fly to help researchers make an informed decision in selecting the most relevant paper son a topic from a combination of archives. (Jacsó 2004).

Over the past two years, Citebase has increased the scope and sources of its full-text content. Previously relying primarily on arXiv, Cogprints, and BioMed Central, as of February 2006, Citebase also harvests OAI metadata associated with full-text documents from 13 additional sources spanning eight countries and representing publisher- and author-based article archives, institutional repositories, departmental archives, national research institutes, international disciplinary archives and university collaborative research teams.

Table 14: Citebase's Data Providers and Record Counts (February 21, 2006)

arXiv.org	357,010
PubMed Central	25,381
BioMed Central	22,339
University of Southampton Eprints Repository	9,137
University of Minho (Portugal) Institutional Repository	3,670
Indian Institute of Science Institutional Repository	3,478
E-LIS: Eprints in Library & Information Science	3,476
Cogprints	2,641
Electronics & Computer Science (ECS) Repository, Southampton	2,561
Lund University (Sweden) Institutional Repository	954
ECS Conference Papers Repository, Southampton	505
University of Otago (New Zealand) School of Business Repository	215
Advanced Knowledge Technologies Repository, Universities of Aberdeen, Edinburgh, Open University, Sheffield and Southampton.	207
National Institute of Agronomic Research (INRA, France) Repository	54
INRA Animal Physiology Repository	16
Organic Eprints Archive, Danish Research Centre for Organic Farming & other international partners	4
	431,648

Source: Citebase/HELP/ Information for reviewers/librarians: http://www.citebase.org/help/info_press.php.

As of February 2006, the database contained more than 430,000 articles, 12.7 million references (of which 2.9 million are linked to the full-text), and approximately 311,000 authors, up nearly 20 percent since July 2005.[106]

Citebase reports an average of 7,000 users on a daily basis. Extensive usage statistics are available from 2002 to present.

---

[106] July 2005 statistics are from Hardy et al. (2005, 55, 56).

Table 15: Comparative Citebase Statistics: July 2005 and February 2006

	Jul-05	Feb-06	Average Monthly Growth	Percent Growth in 7 Months
Number of articles	370,000	431,648	8,807	16.7%
Number of references	10,000,000	12,739,904	391,415	27.4%
Number of references linked to full text	2,500,000	2,937,303	62,472	17.5%
Named authors	260,000	311,021	7,289	19.6%

Table 16: Citebase Usage Statistics

When:	Monthly history Days of month Days of week Hours
Who:	Countries Full list Hosts Full list Last visit Unresolved IP Address Robots/Spiders visitors Full list Last visit
Navigation:	Visits duration File type Viewed Full list Entry Exit Operating Systems Versions Unknown Browsers Versions Unknown
Referrers:	Origin Referring search engines Referring sites Search Search Keyphrases Search Keywords
Others:	Miscellaneous HTTP Status codes Pages not found

Source: http://www.citebase.org/awstats/

Users can search Citebase in three ways: by metadata (i.e. author, title/abstract keywords, publication title, and the date the article was created), citation or OAI identifier. The metadata search engine provides links to abstract/citations pages or cached PDF files (when available). Results are returned in user-specified descending or ascending order according to one of eight rankings:

Search score—relevance rank
Citations by paper
Citations by author
Citation by year
Date created
Date updated
Hits (Web downloads) by paper
Hits (Web downloads) by author, or
By two additional experimental ranks: Hub Score and Authority Score.

Citebase offers ample warnings about how to interpret its coverage and capabilities noting especially that author "hits" are based:

- *only on those citing and cited papers that their authors have already archived in the source eprint archives,*
- *only on those of the cited papers that can currently be successfully linked,*

- *and, for arXiv, for now, on the usage/hit data for its UK-site only.*[107]

In respect to full-text downloads, as of early 2006 they are limited to arXiv (from 1999 to present and UK-site only), Southampton EPrints (from March 2005 with some weeks missing in April 2005) and Southampton ECS repository. Clicking on download statistics generates pie charts and tables indicating when and where the most recent 3,000 full-text downloads occurred (available on an experimental basis for the UK arXiv service only). As of February 2006, nearly 5.5 million full-text articles had been downloaded from Citebase.

Each result is linked to a page of citation tools that provides a graph of the article's citation/hit history; lists all the articles cited by the article (with links out to Google Scholar for each article); identifies the top five articles citing this article (with option to view all articles citing it); and the top five most co-cited articles with this article (with option to view all co-cited articles). The "Correlation Generator" (CG) is a unique tool that provides graphs (or tables) of the correlation between citation impact and usage impact ("hits") from either the UK arXiv.org file or a subset of NASA's Astrophysics Data Service (ADS).[108] In effect, the CG forecasts future citation rates based on Web usage. Southampton researchers posit a positive correlation between initial downloads (i.e. derived from preprints in OA archives) and later citations, suggesting that early Web usage statistics can serve as predictors of later citation impact (Brody, Harnad and Carr 2005).

Steve Hitchcock offers commentary on various studies about the effect of open access and downloads ("hits") on citation impact in a companion (linked) bibliography. Launched in September 2004 in conjunction with Citebase's umbrella initiative "OpCit," this selective bibliography focuses on the relationship between impact and access. Hitchcock estimates that only 20 percent of research articles are published OA despite a growing body of literature offering preliminary persuasive evidence of its positive effect. One section of the bibliography covers the correlation between research assessment rankings and citations (referred to as "the financial imperative.") Although "it does not attempt to cover citation impact, or other related topics such as open access, more generally," Hitchcock includes influential papers as starting points for wider study.[109]

Table 17: Comparison of CiteSeer and Citebase Advantages and Disadvantages

CiteSeer http://citeseer.ist.psu.edu
Advantages [Hardy et al. 2005, 55]

---

[107] Available from HELP: http://www.citebase.org/help/.

[108] These are respectively available from http://www.citebase.org/analysis/correlation.php and http://www.citebase.org/analysis/correlation.php?database=ads

[109] 'The Effect of Open Access and Downloads ("hits") on Citation Impact: A Bibliography' maintained by Steve Hitchcock is available from http://opcit.eprints.org/oacitation-biblio.html .

- Completely autonomous, does not require manual labor.
- Not limited to pre-selected journals or publication delays.
- Searches are based on the context of citations.
- As well as journal articles CiteSeer includes pre-prints, conference proceedings and technical reports.
- User feedback provided on each article (Mathews 2004).
- Can receive email notification of new citations to papers of interest (Lawrence et al. 1999).

Disadvantages [Hardy et al. 2005, 55]

- Does not cover journals that are not available online (Mathews 2004).
- System cannot always distinguish sub-fields (e.g., authors with the same name) (Mathews 2004).

Citebase http://citebase.eprints.org

Advantages [Hardy et al.. 2005, 57]

- Autonomous indexing.
- Easy to use interface.
- Allows users to select the criterion for ranking results.
- Users can rank results by the number of "hits," a measure of the number of downloads and therefore a rough measure of the usage of a paper (Hitchcock et al. 2002).
- Records include informative citation and impact statistics and co-citation analysis with the generation of customized citation/impact charts (Jacso 2004d).
- Additional tools: Graphs of article's citation/hit history, list of top 5 articles citing an article (with a link to all articles citing this article), top 5 articles co-cited with this article (with a link to all articles co-cited with this article) (Hitchcock et al. 2002).

Disadvantages [Hardy et al. 2005, 58]

- Requires better explanations and guidance for first-time users.
- Lacks coverage of a wider range of disciplines.

Source: Compiled from Hardy et al. 2005.

In the intervening two and a half years since the original DLF report appeared, a variety of studies, surveys, and conferences have explored the impact of disciplinary differences on both the use of digital resources and the preferred means of disseminating research results. The "JISC Disciplinary Differences Report" reviews the recent literature and surveys the scholarly communications habits and preferences of 780 academics, representing a wide variety of institutions and departments in the UK (Sparks 2005). One of the key findings substantiates what is already widely known: *the importance of journal articles for the medical and biological sciences; the importance of e-prints (pre and post) in the physical sciences and engineering; the broader mix in the social sciences and the particular importance of books in languages and area studies.*

The survey also corroborates differences in patterns of collaboration and communication, namely that *'harder' disciplines were more likely to collaborate in the research process, and be prepared to use less formal methods to disseminate results, while 'softer' ones were more likely to communicate work-in-progress informally but rely on more formal means of dissemination*. While the survey found *a high level of awareness of current debates about open access across the board*, it also reported that *the overwhelming majority of researchers in all disciplines do not know if their university has an institutional repository*. It comes as no

surprise that physical scientists (44 percent) are most likely to deposit their work in subject archives whereas academics in the arts and humanities are least likely. The majority of this cohort, across all disciplines favored the mandating of self-archiving by research funding agencies.

4.2.11   Current Issues and Future Directions

These resources are at the nexus of key debates about the role and function of different stakeholders in the lifecycle of scholarly information. Authors, researchers, universities, public research funding agencies, publishers, libraries, and vendors are all seeking to reformulate their responsibilities and contributions in view of new modes of creating, organizing, disseminating, and preserving scholarship. Moreover, as evident from the review above, these matters are increasingly played out in highly visible arenas involving national and international advocacy campaigns, policy development, and legislative initiatives. Four inter-related issues come to the foreground:

- The future of the self-archiving movement
- Usage, citation analysis and research impact via Web-based interchanges
- The interplay between disciplinary archives and institutional repositories
- Economic models for Open Access

Each of these is explored further below.

The Future of the Self-archiving Movement

 "Ours is just to deposit and die, not to post endlessly reasoning why…"
        Stevan Harnad, JISC-Repositories listserv, March 16, 2006

Although uptake of self-archiving is on the rise, a survey of author self-archiving habits (N= 1,296) conducted in the last quarter of 2004 found that posting articles at personal Web sites was the most frequent method of publicizing one's work (Swan and Brown 2005). Of the 49 percent who deposited their work, 20 percent used IRs and only 12 percent subject archives. According to another UK study, the vast majority of researchers (N=780) did not know if their university had an IR or not (Sparks 2005). Use of IRs and subject archives for self-archiving varies by discipline, with greatest adoption by physical scientists (Ibid). Swan and Brown report that the vast majority of researchers (81 percent) would willingly comply with a self-archiving mandate by their employer or funding agency. In the absence of such mandates, however, studies in the UK and Australia conclude that only an estimated 15 percent of researchers would voluntarily self-archive their papers.

To increase content in IRs, proponents have discovered ways to align self-archiving more closely with the regular work habits and needs of authors, making it a more

valuable activity serving multiple purposes (Foster and Gibbons 2005). As described earlier in this report, JISC is working with EPrints.org and DSpace to mesh IR submission workflows with the UK's Research Assessment Exercise. The Netherlands has been at the forefront of devising innovative means to encourage self-archiving. "Cream of Science" program showcases the work of prominent Dutch scholars (http://www.creamofscience.org/) and its off-shoot "Promise of Science" aims to bring in ETDs from young scholars.

The actual number of institutional self-archiving mandates is slim at present, although the situation could change rapidly in the next year. At present, four universities in as many countries (Australia, Portugal, Switzerland, and the UK) and two research institutes (CERN and the National Institute of Technology, Rourkela) in Switzerland and India require self-archiving.[110] Policy commitments to OA are more prevalent; three countries—the U.S. (NIH), Germany (German Research Foundation) and Finland (Ministry of Education)—have moved national-level OA policies from proposal to adoption.[111] An important report released by the European Commission in 2006 calls for "guaranteed public access to publicly-funded research, at the time of publication and also long-term" (Dewatripont et al. 2006). Most importantly, in early May, Senators John Cornyn (R-TX) and Joe Lieberman (D-CT) introduced the Federal Research Public Access Act of 2006 (FRPAA) in the U.S. Senate (Bill 2695). According to Suber:

> CURES [described above], FRPAA will mandate OA and limit embargoes to six months. Unlike CURES, it will not be limited to medical research and will not mandate deposit in a central repository. It will apply to all federal funding agencies above a certain size. It instructs each agency to develop its own policy, under certain guidelines laid down in the bill. Some of those agencies might choose to launch central repositories but others might choose to mandate deposit (for example) in the author's institutional repository. Finally, while CURES was mostly about translating fundamental medical research into therapies, with a small but important provision on OA, FRPAA is all about OA. (Suber, SPARC Open Access Newsletter, issue #97, May 2, 2006, http://www.earlham.edu/~peters/fos/newsletter/05-02-06.htm#frpaa.)

The Association of Research Libraries, Association of College & Research Libraries, Association of Health Science Libraries, and SPARC have all endorsed FRPPA. No matter what its fate, FRPPA appears likely to stimulate other countries to push forward with national-level policy adoption.

---

[110] Self-archiving policies and mandates are recorded at the EPrints.org Web site: http://www.eprints.org/events/berlin3/outcomes.html.

[111] Institutional policy commitments are recorded by EPrints.org as noted above. In addition the Berlin Declaration's Web site tracks hem: http://www.eprints.org/events/berlin3/outcomes.html. See also Suber's Open Access News entry of May 23, 2006.

While the legislation is under debate, it is still instructive to review the findings of study assessing authors' understanding of and compliance with the current version of NIH's public access policy (Hutchings and Levin 2006). The survey, conducted on behalf of the Publishers Research Consortium, gives valuable insights into researchers' attitudes and clues about ways to increase compliance. The author reports overall "high awareness but low understanding of the benefits" of the public access policy. It suggests that authors need more details about how the process works, including:

> Whose responsibility it is.
> When it should be done.
> What version is submitted.
> Where it is submitted.
> Where it will appear.
> When it will appear.

In the meantime, some publishers have taken direct control of archiving articles affected by the NIH public access policy, thus alleviating authors of the burden but also usually forestalling deposit until the outermost deadline.

As of this writing, it is safe to conclude that a combination of social, political, cultural and economic factors will affect the future of self-archiving. Obviously national legislation mandating OA deposit of publicly-funded research, as proposed by FRPPA would have a far-reaching effect. In the absence of mandates, self-archiving must become a meaningful activity in its own right, most importantly by demonstrating how it increases visibility and impact of an author's work or brings other added value to busy scholar's work routines. Also hanging in the balance are the role of journal publishers and evaluation of new OA economic models (Walters 2006).

Usage, citation analysis and research impact via Web-based interchange

"Collections principle 6: A good collection has mechanisms to supply usage data and other data that allows standardized measures of usefulness to be recorded."
NISO, *A Framework of Guidance for Building Good Digital Collections*, 2004

As scholarship transitions to the Web, understanding what data needs to be collected and what types of analyses are useful to different disciplines becomes an essential undertaking. Analyzing use (views/downloads), let alone its impact on citations, in an open access environment is a complex affair (Hitchcock 2004-present). Project COUNTER (http://www.projectcounter.org/)[112] and its recent harvesting off-spring,

---

[112] *COUNTER (Counting Online Usage of NeTworked Electronic Resources) is a multi-agency initiative whose objective is to develop a set of internationally accepted, extendible Codes of Practice that will allow the usage of online information products and services to be measured more consistently. Release 2 of the COUNTER Code of Practice for journals and databases was published in April 2005 and is now widely*

SUSHI (Standardized Usage Statistics Harvesting Initiative),[113] are proving effective as a means to gather e-journal usage statistics, and they are in the initial stages of establishing Codes or Practice for collecting usage statistics from IRs repositories.

It is imperative to develop agreed-upon practices for producing article-level statistics from IRs and OA aggregators in the UK because its national Research Assessment Exercise, (http://www.rae.ac.uk/) will move to a metrics-based approach to assessing research quality and allocating "quality-related" public funding after 2008 (UK, HM Treasury 2006). Interoperable Repository Statistics (IRS, http://irs.eprints.org/about.html) is an international effort, led by Southampton University (UK), University of Tasmania (Australia), Long Island University (USA), and Key Perspectives Ltd (UK), under the sponsorship of JISC. IRS complements Project COUNTER and is investigating a coordinated approach to gather and share OAI statistics. IRS, which runs from 2005-2007, has established an international consultative committee that includes principal investigators from various projects reviewed in this report (e.g. CDS, Citebase, DOAJ, OAIster, PubMed Central, SHERPA)

IRS expects to "build generic collection and distribution software for all IRs," and to launch a "pilot statistics analysis service modeled as an OAI service provider." Its principal deliverables are:

- An API for gathering download data implemented for common IR platforms; and
- A set of agreed standards defining the basis for measuring and reporting usage of materials deposited in IRs and aggregated with data from other sources where such materials can be found. (http://irs.eprints.org/)

If successful, IRS will help to overcome some of the challenges noted by OAI services in this section of the report, including scaling, stability, moving from prototype to production services, and most importantly sharing impact values with archives that serve the same documents.

The interplay between disciplinary archives and institutional repositories

The OAI protocol facilitates interoperability across heterogeneous repositories so in the long-run the distinction between archiving in centralized subject-based repositories versus depositing research in dispersed institutional repositories may become irrelevant.

---

implemented. COUNTER is actively supported by the international community of librarians and publishers, and by their professional organizations. (Cited from http://www.niso.org/news/releases/pr-Stats-COUNTER-5-06.html).
[113] For information about SUSHI's Web service harvesting protocol refer to http://www.niso.org/committees/SUSHI/SUSHI_story.html ,.

Presently, in certain arenas, there is a creative tension between these two strategic directions.

Institutional repositories have taken flight since the 2003 DLF survey appeared. By late May 2006, institutional or departmental repositories comprise just about half of the total listings (341 of 686) in the Registry of Open Archives Repositories (ROAR). They are deployed in nearly 40 countries, with more than 100 implementations in the United States alone. (This is a conservative estimate given the voluntary nature of IR registries.) Their low average number of records, just over 3,000, belies their growing influence on campuses worldwide. Nevertheless, it is too soon to tell if they will become the preferred vehicle for depositing and disseminating research output, gaining precedence over discipline-based archives.

As noted earlier, Warner speculates that despite arXiv's overwhelming success, in the long-run IRs may prove more sustainable than subject-focused repositories, which are often dependent on funding from external sources, financially-strapped learned societies or public benevolence on the part of the host institution. Using the field of library and information science (LIS) as an example, Coleman and Roback (2005) argues contrariwise that not all institutions can afford to set up IRs and that subject-based repositories, such as dLIST (Digital Library for Information Science & Technology) and its parent aggregator, DL-Harvest (http://dlharvest.sir.arizona.edu/), may realize economies of scale and have a positive impact on LIS scholarly communication. However, if subject-based aggregators are to "bridge islands of disparities" and achieve their potential, it will require coordinated and strategic planning within the LIS community.

In the field of economics, a high profile pan-European effort is underway to create a disciplinary network that provides integrated access to quality economics resources, drawing on submissions deposited in dispersed IRs. "Nereus" is a consortium of 16 university and institutional libraries with leading economics research ratings in eight European countries, developed in collaboration with researchers (http://www.nereus4economics.info/). According to its developers, "a cornerstone for Nereus is Economists Online (EO)," which "aims to increase the usability, accessibility and visibility of European economics research by digitizing, organizing, archiving and disseminating the complete academic output of some of Europe's leading economists, with full text access as key. EO is building an integrated open access showcase of Europe's top economics researchers based on IRs" (Proudman 2006). Nereus's content will be made available through existing subject search engines and aggregations such as SSRN: Social Science Research Network (http://www.ssrn.com/) and RePEc: Research Papers in Economics (http://repec.org/).

In the UK, JISC has funded a demonstration project (2005-07) to bridge institutional and disciplinary-based repositories. Known as CLADDIER (Citation, Location, And

Deposition in Discipline and Institutional Repositories), it links publications held in two premiere IRs—the University of Southampton and CCLRC (the UK's multidisciplinary research organization)—with data held by the discipline-based British Atmospheric Data Centre (BADC). The goal is to create a system that will enable environmental scientists "to move seamlessly from information discovery (location), through acquisition to deposition of new material, with all the digital objects correctly identified and cited." Experience gained through CLADDIER will be applicable to relationships between other discipline-based repositories and IRs.

Economic models of Open Access[114]

Scholarly publishing—through informal and informal mechanisms—is now in a transitional phase with many unknowns. Willinsky (2006) identifies "ten flavors of open access to journal articles" along with their affiliated "economic models;" he distinguishes the following types of open access (some of which defy strict definitions of OA) and examples:

- Home Page
  Ex: http://www.econ.ucsb.edu/~tedb/
- E-print archive
  Ex: http://arXiv.org/
- Author fee
  Ex: BioMed Central
- Subsidized
  Ex: *First Monday*
- Dual mode
  Ex: *Journal of Postgraduate Medicine*
- Delayed
  Ex: *New England Journal of Medicine*
- Partial
  Ex: *Lancet*
- Per capita
  Ex: HINARI
- Indexing (OA to bibliographic information and/or abstracts, often with pay per view for full text of articles)

---

[114] This is a highly contested issue and well-beyond the scope of this report. For a recent summary of influential studies and different interpretations of economic implications of OA journal publishing, refer to the June 2006 issue of SPARC Open Access News (SOAN) written by Peter Suber, "Good Facts, Bad Predictions" (http://www.earlham.edu/~peters/fos/newsletter/06-02-06.htm#facts). For opposing views refer to the Liblicense-L (Phil Davis). For other studies refer to the Association for Learned and Professional Society Publishers (ALPSP), http://www.alpsp.org/default.htm.

> Ex: ScienceDirect
- Cooperative (members institutions contribute to support OA journals and development of publishing resources)
  Ex: German Academic Publishers (Excerpted from Willinsky 2006, 212-213)

How various economic models play out in terms of aggregations of scholarly information is also an open question. Hailed as the possible "face of the future in online publishing," ResearchNow (http://researchnow.bepress.com/), an aggregation of scholarly materials is featured on the "Best Reference 2005" list by *Library Journal* (LaGuardia 2006, Coutts and LaGuardia 2006). Drawing from three sources—Berkeley Electronic Press (bepress) peer-reviewed journals; contents from participating institutional and subject-based repositories; and items posted directly to the portal via the ResearchNow Upload Utility—this scholarly database is offered in two versions: Open Access and Full Access. In this case, OA is really "quasi-open access," as it offers a combination of restricted views of bepress journals along with unrestricted access to other materials. In the Full Access model, offered at an estimated subscription price of $5,470 per year, all journals and repository materials are available (LaGuardia 2006). Receiving a 5-star rating for its pricing from *The Charleston Adviser*, as of May 2006 ResearchNow boasts more than 100,000 documents and 4-plus million downloads in the past year. Its contents integrate with the e-learning platform, Blackboard and it offers news alerts via RSS feeds. Moreover, ResearchNow is searchable via an XML gateway developed according to the NISO MXG (Metasearch XML Gateway) protocol (discussed in section 4.5). Browsable by subject, in advanced search mode (requires free log-in) results can be displayed as links, XML (Dublin Core DTD) or bibliography export format.

4.3 Pathways to E-Learning in Science and Beyond

This section describes the National Science Digital Library (NSDL) and four related scientific digital libraries, alongside a complementary community of practice in e-learning, MERLOT (Multimedia Educational Resource for Learning and Online Teaching). These services are increasingly anchored and sustained by discipline-based entities as they move from a collection-driven approach to an emphasis on pathways to resources and community participation. Taken together, they serve the full spectrum of "K to gray" learners and educators.

Over the past six years, NSDL has distributed an estimated $125 million dollars in funding to more than 200 projects. While the discussion below concentrates primarily on NSDL's function as an aggregator, harvesting digital resources for discovery via a unified search and retrieval interface, it is important to acknowledge from the outset NSDL's leading role in facilitating research collaboration and engaging stakeholders across public, private, university, K-12, and government sectors in strategic planning for the effective delivery of digital services. NSDL serves a crucial function at the national-

level by re-thinking digital library architectures (Lagoze et al. 2005), developing and promoting best practices (http://oai-best.comm.nsdl.org/), creating generic tools and service applications (http://nsdl.org/resources_for/library_builders/tools.php), conducting research into user needs (California Digital Library 2004, Hanson and Carlson 2005), and advancing techniques in large-project management and participant involvement (Giersch et al. 2004).

The SMETE Open Federation, launched with NSF NSDL Collection and Core Integration funding, includes among its membership more than forty organizations and digital libraries that share the common purpose of advancing digital libraries in science education. The other services discussed in this section are all members of SMETE. NEEDS, BEN, and DLESE are leaders in their respective communities—engineering, biological sciences, and earth science—in building effective digital library services. Although MERLOT's user community is multi-disciplinary, it is included in this section because of its prominent role in science education. It differs from most of the other services under review in this report in two important ways: (1) it is membership-based organization with a formal dues structure that dictates levels of participation and (2) it does not make its metadata freely available for OAI harvesting. MERLOT is particularly known for its peer-review practices and community-developing strategies.

4.3.1 NSDL: National Science Digital Library

Update Table 14: NSDL based on DLF Survey responses, Fall 2005

	The National Science Digital Library (NSDL) http://nsdl.org
ORGANIZATIONAL MODEL	National Science Foundation (NSF)
SUBJECT	Science: STEM (science, technology, engineering, mathematics)
FUNCTION	A digital library of exemplary resource collections and services, organized in support of science education.
PRIMARY AUDIENCE	K-12 teachers, Librarians, NSDL library builders, University faculty
STATUS	Established
SIZE	1.1 million items (265% growth) from 569 collections of which 48 are NSF-funded NSDL collections. 92% of the item-level records are derived from the top 20 collections.
USE	From May to September 2005: unique daily visitors jumped from 8,755 to 11,013; page views increased from 30,106 to 50,440 with 4.65 page views per visit (up from 3.87 in May).
ACCOMPLISHMENTS	1. Improved search service. 2. Improving NSDL data repository using FEDORA. 3. Redeveloped NSDL.org web site that: **--Allows users to self identify by audience** on the homepage in the following categories: K12 Teachers; Librarians; NSDL Library Builders; University Faculty, and; First Time Users.

	--Features periodically updated exhibits, crafted by section editors, for each audience category including: "Top Picks," "Resources of Interest," "Using NSDL," "Research Articles,"Newsfeeds," and an "Events Calendar." **--Provides a one-click connection to browse** by science, technology, engineering and mathematics topics from the homepage. Based on user testing and feedback the new web site design places more emphasis on: **--Active, interactive engagement of users**, Example--"Using NSDL"; **--Being externally focused**, Example--"Newsfeeds"; **--NSDL.org as an educational tool**, Example--"Resources of Interest"; **--Addressing users' educational needs**, Example--"Research Articles"; **--What NSDL has/is**, Example--"Browse by Topic," and; **--What users' want to do or know**, Example--"Ask NSDL."
CHALLENGES	1. Lack of funding to offer more teacher workshops in how to use NSDL through organizations and school districts to increase usage in schools. 2. Great diversity in evaluation methods and tools across 190+ NSDL digital library projects. 3. Lack of a well-funded corporate and foundation outreach program to diversify sustainability options.
TOOLS OR RESOURCES NEEDED	Increased funding for the National Science Foundation particularly EHR (Education & Human Resources).
GOALS OF NEXT GENERATION RESOURCE	In order to increase overall NSDL usage and interactive communications through teacher workshops, professional conferences, and other outreach and communications events and activities, user testing results were analyzed in recreating NSDL.org as a useful educational tool that educators and learners in particular would use repeatedly. Leveraging multiple online and face-to-face interactions is a top priority as repeat users become contributors in a timely and transparent way in the next generation of NSDL.

Since its inception in 2000, the National Science Foundation's Directorate for Education and Human Resources (EHR) has made nearly 220 awards totaling more than $125 million dollars to develop the National Science Digital Library (NSDL).[115] The four major funding streams are defined as follows:

> *Pathways* [replacing "Collections" in FY04] projects are expected to provide stewardship for the content and services needed by major communities of learners.

---

[115] Trying to compile accurate cumulative data about NSF's NSDL funding is surprisingly difficult. The basic sources of data include Lee Zia's annual reports in *D-Lib Magazine* (typically appearing in March) and NSDL annual reports. Although there is award information at the NSDL Web site, it is not kept up-to-date nor does it readily provide dollar-award amounts. To complete FY05 data, the author had to rely directly on the NSF Awards database. The budget data typically reflect the number of awards as opposed to the number of funded projects (awards may be given to different institutions to work on a single project). In addition, Core Integration has numerous sub-contracts but these are not included in award numbers.

Services projects are expected to develop services that support users, resource collection providers, and the Core Integration effort and that enhance the impact, efficiency, and value of the library.

Targeted Research projects are expected to explore specific topics that have immediate applicability to collections, services, and other aspects of the development of the digital library.

Core Integration coordinates and manages the core library, develops the library's central portal and infrastructure, and engages and supports the NSDL community.

Table 18: Summary of NSF NSDL Funding FY2000 through FY2005

	FY2000	FY2001	FY2002	FY2003	FY2004	FY2005	TOTAL
Proposals Submitted	90	109	156	193	144	120	**812**
Total Dollars Requested (in millions)	$59	$64	$92	$110	$126.50	$83	**$534.50**
Funded Budget (in millions)	$13.65	$25.13	$26.76	$22.80	$19.22	$18.00	**$125.56**
Funded Proposals	29	39	55	44	27	22	**216**
Collections (FY2000-03)	13	18	35	22	0	0	**88**
Pathways (FY2004 -)	0	0	0	0	4	9	**13**
Services	9	14	11	11	14	8	**67**
Targeted Research	1	4	6	8	6	2	**27**
Core Integration (CI)	6	3	3	3	3	3	**21**
Subcontracts (part of CI)	0	4	5	5	6	7	**27**

Source: NSDL 2005 annual report; Zia 2001-2005 in *D-Lib Magazine* & email correspondence, March 29-30, 2006.[116]

NSDL's initial emphasis on Collections has shifted over the past two years to configuring and integrating digital resources into sustainable services by anchoring them in established communities of practice thereby "enabling learners to 'connect' or otherwise find pathways to resources appropriate to their needs" (Zia 2006). Collections' funding peaked in 2002 when there were 35 projects, accounting for 68 percent of the total NSDL budget. By 2005, NSDL funding was about equally distributed between Core Integration and the three project tracks, with Services receiving an estimated 28 percent and Pathways, 18 percent of new project funds. To date, Pathways are under development in the biological sciences, physics and astronomy, computational science, middle school teacher resources, materials science, mathematical sciences, engineering,

---

[116] Mick Rhoo's NSDL-CI Evaluation Survey (2006) gives the following figures for the number of awards: Collections, 91; Pathways, 13; Service, 68; Research 27; Core Integration 13—arriving at a total of 212 projects. Report noted in *NSDL's Whiteboard Report*, April 2006, issue 93, available from http://content.nsdl.org/wbr/Issue.php?issue=93.

multimedia resources for the classroom and professional development, and resources and services for community and technical colleges.[117] In FY06 proposals will be accepted for the Pathways track only or "for supplemental funding from existing projects to extend or enhance their services, collections, or targeted research activity so as to enlarge the user audience for NSDL or improve capability for the user."[118] Two projects under review in this report, BEN (BiosciEdNet) and SMETE/NEEDS are exemplars of NSDL Pathways. In addition, DLESE, funded by NSF's Directorate for the Geosciences, serves as an NSDL Earth Science node.

NSDL's 2005 annual report[119] identifies five areas where it is concentrating its efforts to improve education, each with representative project case studies:

- **Evaluation**, the continuous process of measuring the impact of NSDL activities on learning.
  - Case Studies: Teachers' Domain, The BEN Portal, Kinematics Library
- **Classroom Resources**, the nuts-and-bolts work of putting new tools into teachers' hands.
  - Case Studies: Starting Point, TeachEngineering, Instructional Architect
- **Technology**, the massive effort to build and grow a hidden grid that holds digital libraries together.
  - Case Studies: FEDORA Holds Everything, AMSER and CWIS, Searching for Math Formulas (Wolframs Functions)
- **Community Building**, encouraging learning groups to use the NSDL to pursue their questions.
  - Case Studies: Virtual Math Teams, CHEM Collective, Interactive from SHODOR, Environmental Resources Library
- **Informal Learning**, the extension of NSDL resources to libraries, museums, and publications.
  - Case Studies: OCKHAM, *Scientific American* Online, Exploratorium Online

Collections in NSDL

According to NSDL's online collection policy, "NSDL is a collection of other digital library collections." Collections may consist of a single resource or thousands of resources. All NSDL resources are associated with at least one other external collection in order to associate them with a "responsible organization or project." Collections and resources are selected by NSDL Program-funded Pathways and Collections Projects and by the NSDL Director of Collection Development. In addition, collections and resources

---

[117] Descriptions of NSDL Pathways are available from http://nsdl.org/about/?pager=pathways.
[118] Proposal solicitation available from
http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=5487&org=DUE&from=home
[119] NSDL's annual reports are available from http://nsdl.org/publications/index.php?pager=ar .

are recommended by a team of volunteer recommenders (mostly science librarians), NSDL community members, and also the general public. These recommendations are checked against the selection criteria and approved by the Director of Collection Development for inclusion in the NSDL.

There are two broad selection criteria that are intended to be inclusive in order to allow a spectrum of quality and review:

- appropriate to fulfilling the mission of NSDL
- matches the subject scope of NSDL

Users are advised: "NSDL currently contains information about all NSDL funded collection projects, other government funded STEM collections, and other collections associated with universities, private organizations, and companies that fit the subject scope of NSDL."[120]

NSDL collections contain freely available and restricted-use resources. When access is limited, the collection should have open access metadata describing the resources.

As of May 2006, there are 660 collections accepted into NSDL of which 121 have item-level records.[121] Twenty NSDL collections account for 92 percent of the estimated 1.2 million item records. (Lagoze et al 2006a report about their experience in harvesting from 114 NSDL collections via OAI; 37 collections come from only eight providers.) The top twenty data providers range in size from arXiv.org with nearly 340,000 items to DLESE with 7,200 items. The four largest collections also figure among the top twenty in OAIster: arXiv.org, the Office of Scientific and Technical Information (OSTI) OAI Repository, CITIDEL (the Computing & Information Technology Interactive Digital Educational Library), and Wolfram Functions.[122] Of the 88 collection projects funded via NSDL (representing 64 unique collections) from FY00 to FY03 an estimated 75 percent of them have item-level metadata in NSDL. On balance, NSDL-funded collections represent a very small portion of the NSDL content: CITIDEL with more than 100,000 records, followed in size by DLESE with 7,200 items—the remaining circa 46 NSDL-funded collections have 3,000 or fewer records.

---

[120] NSDL's Collection Policy is available from
http://nsdl.org/about/index.php?pager=collection_policy
[121] This represents over 1.2 million resource URLs and 29 registered resource selectors as noted in the *NSDL Whiteboard Report*, issue 93, April 4, 2006, Available from
http://content.nsdl.org/wbr/Issue.php?Issue.php?issue=93
[122] Data about the NSDL collection size obtained in email correspondence with NSDL Director of Collection Development, John Saylor on February 8, 9, 10, and 17, 2006. NSDL's "Collection Registration System" lists accepted and pending collections; it can be filtered by those with or without OAI items: http://crs.nsdl.org/index.php.

Publisher Partnerships in the NSDL[123]

In its work over the past two years, the Core Integration (CI) team has proceeded from the premise that in order for the NSDL to become a resource of choice, used frequently by a broad range of teachers and students on a national scale, it is necessary to engage the interest and participation of the scientific textbook and software publishing community. This community includes both non-profit and for-profit organizations that control a substantial percentage of the high-quality educational science materials currently being produced for teachers and their students.

In this effort, the CI team took steps to engage this community in a collaborative and productive manner, so as to ensure that the NSDL becomes a strong and valued partner rather than a competitor to the traditional science publishing community. Science publishers possess assets that will become critical to the future success of the NSDL, including an efficient and stable mechanism for acquiring and peer-reviewing high quality content from scientists and science teachers; an effective system for editorial development, design, and production of this content; excellent market research and evaluation mechanisms; established models for contracts, licenses, copyright, and intellectual property management; and a reliable system for marketing and sustainability. In addition, many of these publishers work with vendors who provide technical infrastructure and support for schools.

Through its access management and publisher relations efforts, the CI team has established a formal means to engage the science publishing community, including a means to enable controlled access to their content. These activities will ensure that the NSDL reaches its full potential as a functional, valued, and highly used resource, and serves as a model for partnerships with other collaborators in the future.

As of May 2006, NSDL CI has established relationships with 18 science publishers. Many of these have begun to supply metadata for their materials which then appears in the NSDL central portal interface. The publishers include:

- American Mathematical Society
- American Physical Society
- Bedford, Freeman, and Worth
- BioOne
- Blackwell Publishing
- Cambridge University Press (book and journal programs)
- Elsevier Books
- Houghton Mifflin Company
- John Wiley and Sons

---

[123] This text was contributed directly by NSDL's Core Integration team in May, 2006.

- McGraw-Hill
- National Academy Press
- Nature Publishing Group
- Oxford University Press (book and journal programs)
- Pearson Education
- Scientific American
- Springer Science+Business Media
- Tom Snyder Productions (software division of Scholastic)
- Tool Factory (educational software)

Navigating NSDL

Users can browse collections alphabetically by title or by an expandable subject tree (branching out from Education, Health, Mathematics, Science, Social Studies, and Technology). Collections can be also identified through the interactive visual view of "NSDL At a Glance" tool, organized by topics from The Gateway to Educational Materials (GEM) subject scheme.

Figure 25: Screenshot of NSDL at a Glance: Astronomy Interactive Visual View of Relevant Collections
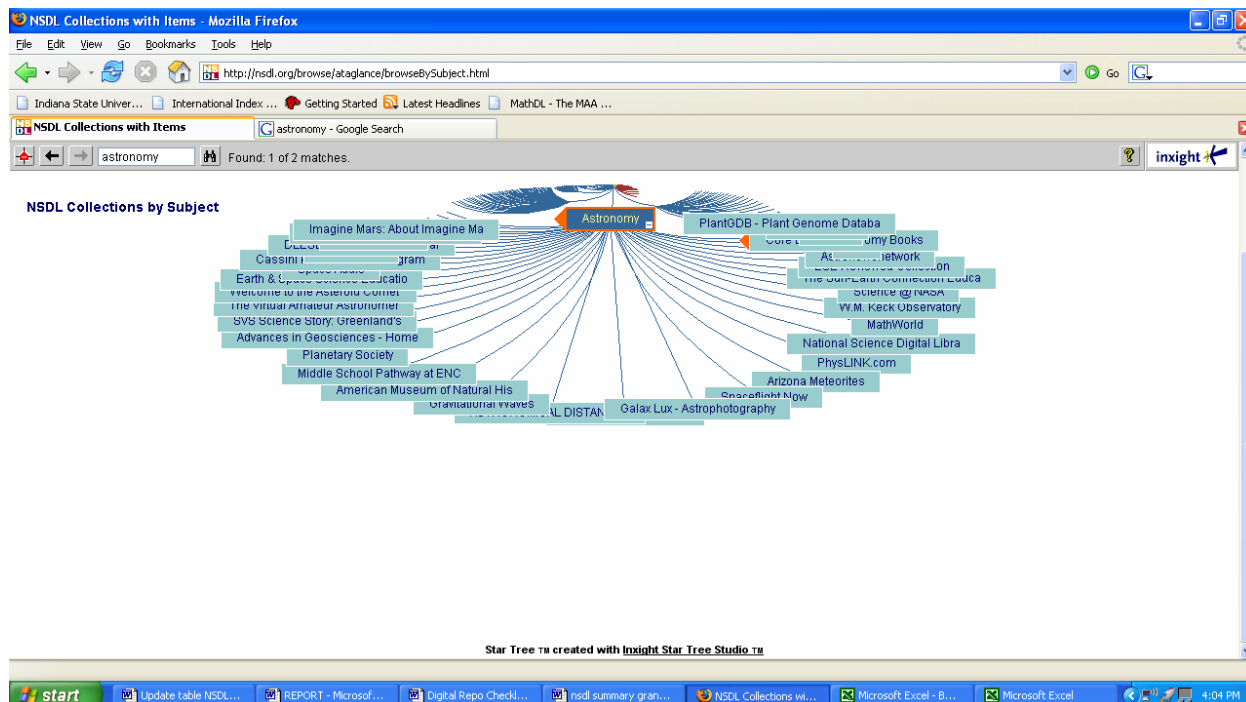
Source: http://nsdl.org/browse/ataglance/browseBySubject.html (February 2006)

Since the 2003, NSDL has developed access point to its content by audience: K12 Teachers, Librarians, NSDL Community, University Faculty, and First Time Users.

Table 19 summarizes the widely varying results retrieved in a search for resources relevant to University Faculty about Astronomy. Browsing by topic identifies 68 collections relevant to Astronomy. A keyword search retrieves more than 11,000 resources. The University Faculty portal contains one "top pick" relevant to Astronomy. A search of the Virtual Reference Desk, AskNSDL question-and-answer archives and resources (requires registration and log-in) compiled by an NSDL reference desk librarian, locates 12 collections relevant to Astronomy (but the list does not include the Physics and Astronomy Pathway found through the University Faculty portal). In addition to blogs, Web sites and other types of resources, AskNSDL has 68 archived questions from users related to Astronomy.

Table 19: Astronomy Results from Various Portals, Pathways, and Navigational Features

Browse by Topic: 68 resources
Search by Keyword: 11,091 resources

University Faculty Portal
• **Top Picks:** 1 (of 14)
ComPADRE
NSDL PHYSICS AND ASTRONOMY PATHWAY
To Physics and Astronomy Education Resources
Through a partnership of authors and organizations ComPADRE acts as a steward for the educational resources used by broad communities in physics and astronomy by creating and sustaining a network of collections that provide learning resources and interactive learning environments. ComPADRE resources positively influence physics and astronomy students and their teachers in both individual and collaborative settings.
• **Resources of Interest:** 0 (of 8)
• **Using NSDL:** 0 (Although at least 2 of 5 resources featured are of potential interest)
    1.   Sunshine Applet
This Java applet shows sun exposure and intensity for any latitude and longitude, and any date during the current year. The times of most intense and dangerous sunshine are given through a chart and global map, as well as a graph indicating the current location of the Sun in terms of strength. There is also an indication of sunrise, sunset, and Sun culmination.
    2. ATHENA Mars Exploration Rovers
Cornell University, NASA's Jet Propulsion Lab, and Bill Nye present information on the Mars Athena Exploration Rovers. Mission updates from Athena Principal Investigator Steve Squyres, technical briefings, Images, at-home experiments for kids and lesson plans compliment details of mission goals and payload.
• **Research Articles** (0 of 7)
• **News Feeds:** 0
• **Events Calendar:** 0

AskNSDL: **Home: Science: Astronomy: Resources**
• **Blogs, Feeds, Podcasts:** 5 resources
• **FAQs:** 9 resources
• **Suggested Web Sites:** 10 resources
• **Archived NSDL Scout Reports:** 1 resource
• **NSDL Collections:** 12 resources
    Astronewsnetwork
    Core List of Astronomy Books
    Exploratorium. Ten Cool Sites: Astronomy
    ibiblio
    NTRS: NASA Technical Report Server
    PhysLINK.com - Reference and Education - Physics, Astronomy and Engineering
    SEGway: The Science Education Gateway
    Smithsonian Institution
    Spaceflight now - The leading source for online space news
    The Parallax Project
    The Sun-Earth Connection Education Forum

Virtual telescopes in education (VTIE)
- **Other "Ask an Expert" Archives:** 1 resource
- **Educator Resources (lesson plans):** 1 resource

AskNSDL: Home: Science: Astronomy: Archived Questions: 68 Questions & Answers

Source: http://www.nsdl.org/ (February 2006)

Given the diversity of these sample results, users should be encouraged to experiment with different search, browse and navigational functions to see which best suit their needs.

Search Features

In 2003, NSDL offered both simple and advanced search features. Simple search relied on keywords with the ability to limit by Type of Resource (Collections, Items, News, Exhibits, Collections with reviews, Items with reviews) or by Resource Format (Text, Image, Audio, Video, Interactive Resource, Data), whereas advanced searches allowed Boolean commands limited to keyword anywhere, keyword in content, title, author/creator/contributor, subject and format/genre. In response to user feedback, NSDL simplified its approach and now offers a single search box for keywords with the option to limit the search by Resource Format (same as above) or Grade Level (Graduate, College, High school, Intermediate elementary, Middle school, Primary elementary). In spring 2006 NSDL added an option to *Search resources* (i.e. educational resources) or *Search NSDL.org* (i.e. NSDL community sites or the NSDL.org site). As of this writing, these labels are under review and additional information describing the options will be added once approved.

Figure 26: Screenshot of NSDL's Search Page
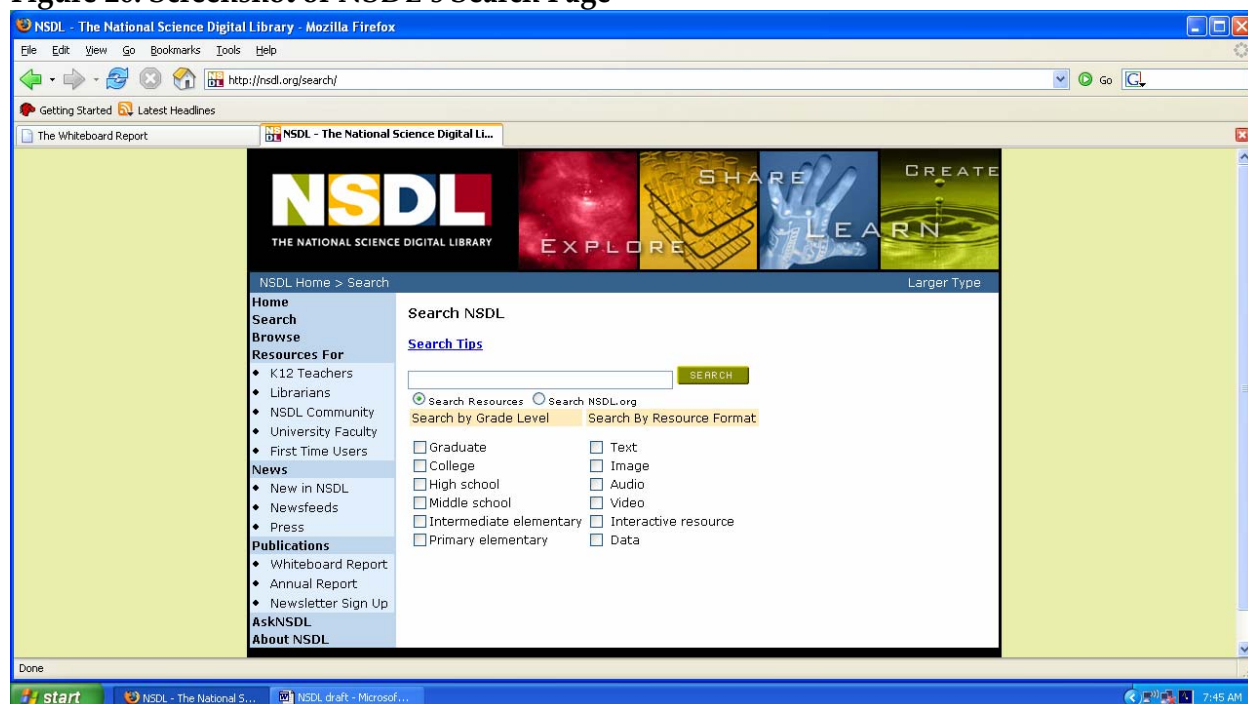
Source: http://nsdl.org/search/ (May 2006)

Search Tips explain that searches are not case sensitive and that quotation marks should be used around phrases. Boolean commands are no longer available, nor is there any explanation whether or not there is automatic ANDing of search terms (a common feature of most general search engines) or truncation (or other wildcard functions).

Results are returned ten per page with brief annotations and links to "View all related information" (provides item and collection-level metadata, as available) and "Include/Exclude results like this" (enables filtering by collection). Users can navigate to previous or next pages but cannot sort results or jump to different page results. When the revised Web site went live in late October 2005, a standard feature was added so all NSDL.org pages can be emailed via the "Email this page" link in the footer. However, there are no post-processing features to save or export results by other means. When a search does not produce any results, users are advised to consult the search tips, browse the collections, or check back as NSDL collections continue to grow. In March 2006, NSDL implemented "Did you mean" spelling suggestions. A search for "crystalography," suggests the corrected spelling, "Did you mean," crystallography.

Search Results

When conducted in late January 2006, a sample search for the keyword <crystallography> produced curious results (Table 20). Without deploying any search delimiters, the basic term query returned 633 results. Filtering the results to exclude the

collection of the first retrieved item, reduced the result set to 616 resources. When the link "search within this collection" (e.g., DLESE) is used, the results increased dramatically to 7,175. Beginning anew with the search term, <crystallography,> but limiting by grade level, produced a wide range of results, with 20,373 hits at the high school level.[124] Given the proviso that not all resources contain format metadata and, therefore, relevant results may be excluded, it was alarming to retrieve much higher and wildly different returns when the format delimiter was invoked—e.g., over 700,000 texts and 34,000 images pertaining to crystallography, when the keyword search retrieves 633 resources. Based on Brogan's query of early February 2006, it became apparent that NSDL was combining all keyword appearances (i.e. through OR operators) rather requiring the presence of both words (i.e. through AND operators). NSDL modified its newly implemented search interface and corrected these errors on the production site in mid-February 2006. The results of the identical search conducted after the modification are dramatically different.

Table 20: Search Results for <crystallography> with and without delimiters

SEARCH QUERY	RESULTS: January 31, 2006	RESULTS: March 21, 2006
KEYWORD: crystallography 1st Item Retrieved: crystallography According to the annotation: *This site is a link from a mineralogy database hosted by webmineral.com.* "View all related information" indicates that this item is derived from the DLESE (Digital Library for Earth Science Education) Collection. • Include/Exclude results like this Exclude This Collection (e.g., DLESE) • Search within this collection (e.g., DLESE)	• 633 • 616 • 7,175	• 824 • 802 • 22
SEARCH BY GRADE LEVEL: crystallography • Graduate • College • High school • Middle school • Intermediate elementary • Primary elementary	 • 5,259 • 6,520 • 20,373 • 12,159 • 2,819 • 6,198	 • 12 • 6 • 10 • 2 • 1 • 0
SEARCH BY FORMAT: crystallography • Text	 • 702,783	 • 239

---

[124] McCown et al. (2005) recommended that NSDL provide an advanced search capability and the ability to target by grade level. It would be useful to conduct a more thorough study now that NSDL has addressed many of these recommendations and has implemented a new search service. The McGown study was conducted before NSDL began to use the Cornell-based search service in mid-2005. The original study evaluated search results for pedagogical resources in NSDL and Google, finding that a significant portion of NSDL's resources had inaccessible content (for a variety of reasons), that in four of six subject areas Google significantly outperformed the NSDL, and that Google's precision was superior to NSDL's.

• Image	• 34,167	• 16
• Audio	• 1,350	• 0
• Video	• 5,743	• 0
• Interactive resource	• 9,463	• 4
• Data	• 771	• 1

However, as NSDL officials explain, determining how "boosting and filtering" occurs is not entirely straightforward; in the many cases where the data provider or collection does not provide resource-type information in their metadata, relevant results may be lost from the search and the results are narrow. Even so there are still other problems apparent in this new sample search. The 12 results for Graduate-level resources contain three apparent duplicate references to Reciprocal Net. Users have to link to another screen to find out if all three are from the same source or not. Two seem identical (despite different NSDL OAI identifiers); the third is an article discussing Reciprocal Net that appeared in a NSDL *Whiteboard* report. Moreover, Reciprocal Net is tagged for three grade levels "graduate, undergraduate, grades 10-12" yet only shows up in the "Graduate" search. In early April 2006, NSDL reinstated the collection icons in search results pages allowing users to see the collection in which the resource resides, which helps to address some of these issues.

NSDL's New Resource-Centric Fedora Architecture[125]

NSDL's conversion to a Fedora repository marks a major transition from a metadata-centric to a resource-centric data model and search service. According to NSDL developers:

> Digital libraries need to distinguish themselves from web search engines in the manner that they add value to web resources. This added value consists of establishing context around those resources, enriching them with new information and relationships that express the usage patterns and knowledge of the library community. The digital library then becomes a context for information collaboration and accumulation – much more than just a place to find information and access it. (Lagoze et al. 2005)

Finding the metadata-based model inadequate, the developers describe "an information network overlay within Fedora, which includes the full functionality of the existing metadata repository, but models relationships, services, and multiple information types within a web-service based application" (Lagoze et al. 2005). More recently, NSDL principals analyzed the many difficulties they have encountered over several years in relying on metadata to build the NSDL. They provide persuasive evidence for the new

---

[125] An Overview of the NSDL Library Architecture (dated 11/10/05) is available from http://nsdl.comm.nsdl.org/docs/nsdl_arch_overview.pdf.

"resource-centric architecture that integrates less structured forms of information, which collectively add value and context to digital resources." As they explain:

> Traditional structured metadata plays a role in such information contextualization. However, it exists as a component of a resource-centric model, rather than being the focus of the information model itself. (Lagoze et al 2006a, 3)

Their discussion goes beyond metadata quality to investigate other issues that add complexity and cost to operating a large-scale metadata aggregation site like the NSDL. For example, they reveal dismal harvesting statistics, citing an overall success rate of 64 percent and a monthly failure rate of 25 to 50 percent. They attribute harvest failures equally to three broad areas:

1. a communications or system failure either at the data provider's server or with the NSDL's OAI harvester
2. OAI protocol violations
3. invalid XML data, XML schema non-compliance, or SML, URL or UTF-8 charactering encoding (Lagoze et al 2006a, 5)

Resolving harvesting failures entails extensive email communication, estimated at 170 messages per provider per year.

The new architecture is intended to model *resources* rather than *metadata* and permit the provision of richer information, including context and less-structured metadata. The infrastructure is also making possible a number of new NSDL applications described by Lagoze and his colleagues:

> *Expert Voices* – a collaborative blogging system that enables such capabilities as Question/Answer discussions and resource recommendations and annotations.
> http://expertvoices.nsdl.org/?css=larger
> *On Ramp* – a system for the distributed creation, editing, and dissemination of content from multiple users and groups in a variety of formats.
> http://nsdl.comm.nsdl.org/docs/nsdl_arch_overview.pdf#search='on%20ramp%20nsdl'
> *Instructional Architect* – a system that enables teachers to identify, choose and design lesson plans, study aids, homework using online learning resources.
> http://ia.usu.edu http://ia.usu.edu

Content Assignment Tool – a tool to align NSDL resources to state and national education standards.[126] http://www.cnlp.org/documents/casaa-web/casaa.html

Sustaining NSDL Collections and Services

Faced with the prospect of diminishing NSF funds, NSDL is increasingly turning its attention to strategies that will sustain its efforts and integrate them into established library services. NSDL's Sustainability Standing Committee Chair, Paul Arthur Berkman outlines four components of NSDL, each requiring its own strategies, if the NSDL is going to survive as a collaborative, coordinated effort where the sum is greater than its parts.

- **Program Sustainability** involves strategies to facilitate long-term collaborations among projects, uses, sponsors, federal agencies and other stakeholders that share in the progress of the NSDL.
- **Project Sustainability** involves the public-private-university-government strategies to support the creation, maintenance and evolution of collections and services in the NSDL.
- **User-Community Sustainability** involves the networking, outreach and engagement strategies that are necessary to grown the community of users, members and sponsors who will support the NSDL into the future.
- **Technical Sustainability** involves coordination among technology developers and the overall program to develop the NSDL in a persistent, functional, and visionary manner.

In 2004 NSDL began to publish "sustainability vignettes" in the *Whiteboard Report* for specified projects. The seven vignettes issued to date represent a range of multi-faceted approaches to continuation.[127] The Math Digital Library, for example, is creating new value-added services in close consultation with members of the Mathematical Association of America (MAA). According to MathDL's vision, new components—for example, MAA Reviews, Classroom Capsules, online MathDL books and meeting and workshop software—would be free to members but non-members would be required to subscribe or pay a usage fee. Similarly, MathDL's Journal of Online Mathematics and its Applications (JOMA) may transition to a member-only benefit, requiring others to pay for access. In brief, MathDL's sustainability plan hinges on a combination of support from MAA and from direct income streams. Another NSDL project, Teacher's Domain, sponsored by WBGH, is seeking "collaborative partnerships and strategic alliances,"

---

[126] Lagoze et al. (2006b, 5) refer to the Content Alignment Tool; whereas the project Web site calls it a Content Assignment Tool.

[127] Five vignettes are accessible via the NSDL's Community Portal for the Sustainability Standing Committee http://sustain.comm.nsdl.org/. Or to retrieve all six reports, search the *Whiteboard Report* <sustainability vignette> from http://nsdl.org/publications/index.php?pager=wbr.

along with the expectation that its courses will become self-funded through licenses to educational institutions and organizations.

The NSDL Sustainability Standing Committee is developing a decision-tree exercise, designed to help principal investigators determine if and how to sustain their NSDL projects. Alternative decision paths branch out from responses to questions about the project's sustainability objectives, its relevance, institutional support, and market opportunities—resulting in recommendations to discontinue the project as unsustainable or to consider open-source community, not-for-profit or for-profit corporation resolutions.

Several other initiatives, addressing user-community and technical sustainability, merit discussion. Effective October 1, 2003, the California Digital Library (CDL) received a two-year NSF grant to develop and enrich the NSDL by determining how to best integrate it into academic library services. In an effort to support the development of NSDL's long-term business plan, the grant provided for a market assessment to determine user needs and expectations of high-quality science online resources. Through focus groups, interviews, and a comparative review of user-specified high-quality science resources (e.g., HighWire, Scirus, PubMed, CiteSeer), CDL market research revealed:

- Limited prior awareness of NSDL; lack of differentiation vs. other government science Web sites (e.g., Science.gov).
  - N.B. In May 2006, NSDL announced that Science.gov had added NSDL to its collection. According to the announcement in *NSDL's Whiteboard Report*: This means that users can search all the science databases and more than 1,800 science Web sites at Science.gov (http://www.science.gov/), plus the 1.1 million records of science, technology, engineering and mathematics education resources at NSDL, with just one click.[128] (See Figure 27 below.)
- Strong resistance to institutional subscription model, especially in current California K-12 funding climate.
- Most participants see more value in the NSDL collection as a classroom teaching aid for K-12.
- Academic libraries see limited value in another Web science portal, but would be willing to consider paying for deep integration with their existing search tools.
- Mixed levels of interest in personalization and publishing tools. (California Digital Library 2004, 1)

---

[128] *NSDL Whiteboard Report*, May 2006, issue 95 available from http://content.nsdl.org/wbr/Issue.php?issue=95.

There is further evidence from this 2006 review of NSDL's functionality (and its new technical infrastructure) that the findings and recommendations of CDL's market assessment are informing NSDL's current development and fund allocation. For example, in April 2006 NSDL Core Integration was awarded a grant in collaboration with Utah State University, and SUNY-Cortland to help teachers learn to design educational activities with NSDL resources that will lead to more teacher-designed and contributed content in NSDL and will also measure the impact of project activities on teaching practice.[129]

CDL's recommendations are annotated below with checkmarks to indicate areas of subsequent progress ("o" indicates not implemented as of mid-May 2006):

- NSDL should provide **free, open access** to its basic collection through a public Web portal that provides basic metasearch features and a browsable subject hierarchy.
    - ✓ Browsable subject hierarchy instated.
- **Improve current NSDL portal** by improving visibility of search, creating browsable subject hierarchy in HTML, and a clear statement of purpose and intended audience.
    - ✓ NSDL has developed five entry points geared towards different audiences.
- **Encourage K-12 classroom use** by providing access to lessons plans, subject guides, and interactive features; consider partnering with established K-12 content providers.
    - ✓ These features are available via the ASKNSDL service.
    - ✓ Content Assignment Tool aligns national and state educational standards to resources.
    - ✓ Several Pathways partners are addressing this, e.g., WBGH's Teacher's Domain, AAAS's Biological Sciences Pathway (via BEN portal), Engineering Pathway (merger of NEEDS and TeachEngineering) and the Middle School Pathway (Ohio State University).
    - ✓ Established partnership with the National Science Teachers Association (NSTA) to deliver 11 *NSDL Online Science* Web seminars through June 2007.
- Explore development of **value-added services for academic libraries**, including:
    - o MARC record export
    - o OpenURL support
    - o Integration with other federated search platforms
    - o Mapping of controlled vocabularies (e.g. MeSH-type thesaurus)

In addition, CDL recommended that NSDL evaluate incorporation of various features suggested by focus group participants. Items with checkmarks have been implemented:

---

[129] See NSF grant #TPC 055440, 2006-09 for details.

- o Citation linking
- o Abstracts
- o "Smart parsing" of search terms (e.g., cell biology > "cell biology")
- o Suggest related terms based on search input
- o Search history
- ✓ Ability to rank search criteria: NSDL removed rank from search results based on user testing. Results are sorted based on rank.
- ✓ Image search tools (e.g., Browse, "NSDL At a Glance")
- ✓ "Search within these results": beta version in place on the development server, not in production.
- ✓ Personalized views of the collection
- ✓ Community features (e.g., discussion forums, listservs, RSS for registered users) (Features list from CDL 2004, 20; annotated with check-marks by author)

A second focus of the grant is to develop a prototype service that integrates NSDL into "foundational science collections managed by libraries" and provide the tools to create different views "customized to the needs of different patrons."[130] As of this writing, the prototype NSDL service integration is not yet available, but CDL is creating a portal for the geosciences (FindIt: Earth Science) that offers users a unified interface to search domain-specific proprietary databases (e.g., GeoRef, Web of Science) alongside OAI-harvested NSDL and DLESE records and items retrieved via targeted Web crawling.[131]

Other major services, for example Science.gov, are starting to integrate NSDL resources into their search capability.

---

[130] Project abstract is available from http://www.cdlib.org/inside/projects/metasearch/nsdl.
[131] See section 4.5.4 of this report for a discussion of CDL's report on resource integration (Christenson and Tennant 2005).

Figure 27: Screenshot of Science.gov and announcement about NSDL

Source: http://www.science.gov/ (May 6, 2006)

Finally, the OCKHAM Initiative (described in section 3.2), led by Emory University and Oregon State University, aims to establish "an extensible framework for networked peer-to-peer interoperation among the NSDL and traditional libraries." To this end, it is developing a suite of tools (middleware) to help integrate NSDL collections and services into traditional library service environments while also creating a current awareness alerting service and a registry to facilitate machine-to-machine and end-user discovery of digital library services. This is vital to the future effective interoperation among existing NSDL collections. In addition, NSDL and DLF are working together to establish and promulgate Best Practices in Shareable Metadata as discussed earlier in this report.

Leveraging Individual Project Activities and External Relationships

This description of NSDL concentrates in large part on its Core Integration activities as an aggregator of STEM collections and services. NSDL, however, makes many other valuable contributions to advancing STEM teaching and learning by leveraging partnerships between individual projects and national partners. This is particularly evident in NSDL's involvement in the promoting educational achievement standards and professional development workshops. To cite one prominent example, the NSDL Achievement Standards Network (ASN), developed with NSDL funding by Jes & Co. (http://www.jesandco.org/), will provide hands-on learning standards systems for every state. The NSDL resource records are part of the State Educational Technology Directors Association (SETDA) 2006 Tool Kit (http://www.setda.org/content.cfm?SectionID=265),

developed in conjunction with the U.S. Department of Education. The initiative includes tools, technologies and best practices that enable states to manage electronic versions of their academic standards, align resources and assessments consistently using open and interoperable methods, and embed standards seamlessly in all manner of learning and assessment systems and systems of accountability.

4.3.2 SMETE: Science, Mathematics, Engineering and Technology Education Digital Library

Update Table 15: SMETE based on DLF Survey responses, Fall 2005

	SMETE Digital Library http://www.smete.org/ **NEEDS: National Engineering Delivery System** http://www.needs.org/needs/
ORGANIZATIONAL MODEL	Open federation, voluntary membership w/ partners and affiliates funded by NSF and other public/private agencies
SUBJECT	Science: science, mathematics, engineering & technology
FUNCTION	Collection of collections and community of communities
PRIMARY AUDIENCE	Educators
STATUS	Established
SIZE	9,500 resources in SMET disciplines including 2,200 engineering resources
USE	Per month: 30,000 page hits
ACCOMPLISHMENTS	1. Interoperability with other digital libraries. 2. Providing digital repository services, e.g., Digital Chemistry, Exploratorium, NCWIT (National Center for Women & Information Technology). 3. Community development with Premier Award and monthly theme pages.
CHALLENGES	1. Expanding community building beyond ASEE (American Society for Engineering Education) audience. 2. Sustainability planning. 3. Quality control of metadata and contents of the learning objects in merger between NEEDS (National Engineering Education Delivery System) and TeachEngineering: Resources for K-12
TOOLS OR RESOURCES NEEDED	1. Push technologies. NSDL On-Ramp. 2. Community building tools, e.g., Threaded discussion forums, Blogs, Newsletters.
GOALS OF NEXT GENERATION RESOURCE	NEEDS will be merging with TeachEngineering to form the new Engineering Pathway to serve the entire engineering education community from K-12 to lifelong learning. SMETE.org will continue to be NEEDS technology platform to provide supports for other online learning projects such as the Mobile Learning project sponsored by HP and CITRIS (Center for Information Technology in the Interest of Society).

The SMETE Open Federation continues as a membership organization launched with NSF NSDL Collection and Core Integration funding whose "primary mission is to establish universal access to academic excellence in SMET education." The Federation has more than forty partners including the American Association for the Advancement of Science (AAAS), the Coalition of Networked Information (CNI), and OCLC as well as other digital libraries dedicated to science education (including all of the services under review in this section) and a dozen universities and corporations. SMETE helps to develop leading-edge technologies to share among its members while also maintaining a collection of premier learning materials.

SMETE collaborated with the Exploratorium, in San Francisco, California, to create the Exploratorium Digital Library, a collection of high-quality teaching resources and activities (http://www.exploratorium.edu/educate/dl.html) that is also integrated into NSDL. SMETE has also provided technology services to other digital libraries including BioSciEdNet (BEN, http://www.bisciednet.org/portal/), MathDL (http://mathdl.maa.org/), and the Digital Chemistry (http://socrates.berkeley.edu/~kubinec/). SMETE resources are cataloged to meet the requirements of the IEEE Learning Object Metadata Standard and SMETE has developed tools to transform local application profiles (e.g., from LON-CAPA, http://www.lon-capa.org/ and the Michigan Teacher Network, http://mtn.merit.edu/) to normalized application profiles. SMETE collaborates with MERLOT on peer reviews.[132]

In addition to supporting search queries by keyword, author/creator, title, and publication date range, the user interface offers various options to limit searches by more than 20 different types of learning resource (e.g., case study, dataset, lesson plan); grade level (primary education to post-graduate and vocational training to professional development); and eight specific collections (e.g., ACM Women in Computing, Math Forum, Michigan Teachers Network, NEEDS). Searches can be restricted to peer-reviewed materials. Search results are returned with briefly annotated entries including a search score. Each result is clearly branded according to its platform (e.g., PC, MAC, Web); cost (e.g., free or $); availability of reviews; and native collection. Registered users can create a profile and save resources in a workspace. User information can be shared to identify other community members with similar interests.

The results' screen provides users with related terms to extend the search as well as the ability to conduct a federated keyword search in partner collections. The partner collections include NSDL, MERLOT, and NEEDS. A technical report available at SMETE

---

[132] For additional details refer to the final NSF aware report, Agogino, A.M. 2004. Enhancing Interoperability of Collections and Services. Final Report, December 2004. NSF Award DUE-0127580. Available from http://best.me.berkeley.edu/%7Eaagogino/papers/Final_Report_SMETE.pdf .

explains its strategy for adopting a SOAP-based SMETE Search API to implement federated searches across heterogeneous collections.[133]

4.3.2.1 NEEDS: National Engineering Education Delivery System

The American Society for Engineering Education in partnership with seven leading engineering schools (e.g., UC-Berkeley, Worcester Polytechnic Institute, Colorado School of Mines) is creating a unified K-gray engineering pathway, under the auspices of NSDL. NEEDS, a digital library for engineering education, will merge with TeachEngineering (Resources for K-12) to establish a single comprehensive portal for engineering. Both NEEDS and TeachEngineering (TE) are highly regarded by their respective communities. Through its annual "Premier Award" courseware competition, NEEDS is a national leader in stimulating and evaluating high-quality engineering courseware targeted for undergraduate teaching. It has translated the award selection criteria into best practices in courseware design, helping to promulgate high standards of excellence. Through the combined expertise of NEEDS and TE, they expect to:

- Significantly and sustainably grow high-quality resources;
- Align the unified curricular materials with appropriate undergraduate and K-12 educational standards;
- Grow the participation of content providers and users;
- Enhance quality control and review protocols for content; and
- Expand gender equity and ethnic diversity components by cataloging and reviewing curricular resources created by female-centric and minority-serving organizations.

As an initial step in developing a unified service based on SMETE's technology platform, NEEDS and TeachEngineering (TE) launched a blog to discuss desirable features for the new pathway. An initial list of tools and services included:

- Browse curriculum (TE)
- Search resources/curriculum by Keyword, Grade Level, Educational Standard (TE), Required Time (TE), Cost (TE), Learning Resource Type (NEEDS), Title (NEEDS), Author (NEEDS), Review (NEEDS), Series (NEEDS), Host Collection (NEEDS), Publication Year (NEEDS)
- Personal workspace (MyTE, NEEDS Workspace)
- Reviews for resources/curriculum
- OAI server that exports NSDL Dublin Core metadata for harvesting
- Recommendation system (NEEDS)
- Web service for search through SOAP (NEEDS)

---

[133] A bibliography of publications and presentations about SMETE/NEEDS is available from http://www.needs.org/needs/public/about_needs/publications/

- Metathesaurus to suggest related search terms (NEEDS)
- RSS feeds of new resources (NEEDS)
- Online cataloging (NEEDS)[134]

4.3.3 BioSciEdNet (BEN) Collaborative

Update Table 16: BEN Collaborative based on DLF Survey responses, Fall 2005

	BiosciEdNet (BEN) Collaborative http://www.biosciednet.org/
ORGANIZATIONAL MODEL	Collaborative sponsored by the American Association for the Advancement of Science and other disciplinary organizations.
SUBJECT	Science: biological sciences
FUNCTION	Portal to digital libraries for teaching and learning in the biological sciences.
PRIMARY AUDIENCE	Educators
STATUS	Established
SIZE	Collaborators increased from 15 to 22 (46.6% growth). Peer-reviewed resources grew from 1,000 to 4,100 (310% growth). Registered users grew to 5,500 and 92% are educators. BEN covers 76 (previously 51) topics in the biological sciences.
USE	Per month: >1.4 million visitors to the BEN portal and collaborator sites. ~6,000 registered users: 91% teach (62% at undergraduate and 19% at high school level).
ACCOMPLISHMENTS	1. Initial development of models for transforming smaller organizations into contributors of resources to digital libraries. 2. Increased the number of peer-reviewed individual biological sciences learning objects or resources. 3. Conducted a BEN User Survey in September 2004, where 515 responses were returned in a 3-week timeframe, representing a 14% return rate.
CHALLENGES	1. Building and supporting a diverse contributor/user base for the digital libraries is one of the most critical issues that BEN faces. Since undergraduate biology is a core course in many colleges and universities and high school biology educators tend to teach 4 to 5 biology classes a day, these educators often have severe constraints on both time and resources. 2. Building digital collections that are inclusive of all educators and students. Biological sciences educators, particularly in high schools and community colleges and regional comprehensive institutions, have student bodies diverse in every respect – learning styles and ability, geography, economics, race, gender, physical disabilities, and experience. 3. Streamlining and lowering the barriers to participation by additional organizations that develop high-quality peer-reviewed bioscience educational materials, but don't have the technology or staff to develop digital library collections from the ground up.
TOOLS OR	Development of a BEN Faculty Campus Representative Program for

[134] Engineering Pathway blog posting of November 17, 2005. Available from
http://dev.needs.org:9006/blojsom/blog/default/

RESOURCES NEEDED	Increasing Contributors and Users of BEN and the NSDL. Establishment of mentor relationships between mature and new BEN Collaborators. Provide software tools for BEN Collaborators.
GOALS OF NEXT GENERATION RESOURCE	1. To increase the number of Collaborators that BEN aggregate resources from 13 to 22. 2. Through mentoring and technical assistance to other organizations, the total number of biological sciences digital libraries developed by members of the BEN Collaborative would increase from 6 to 13. 3. Develop a Faculty Campus Representative Program, including related professional development, materials and a demonstration CD ROM. Through the Faculty Campus Representative Program, 45 college and university faculty members, geographically dispersed around the US, will be prepared to provide campus and community-based workshops and technical assistance in selected areas for an estimated 2,700 prospective contributors to both BEN and the NSDL.

In fall 2005 the BEN Collaborative, led by the American Association for the Advancement of Science (AAAS) with a dozen founding-partner professional societies, received NSF NSDL funding to expand into a Biological Sciences Pathway for educators at the high school and undergraduate levels.[135] Over a four year period, the Pathway funding will enable BEN to increase the number of: collaborators from which it aggregates resources from 13 to 22; digital libraries it helps professional society members to develop from 6 to 13; and cataloged resources in the BEN metadata repository from 4,000 to 27,000 items. With more than 100 professional organizations in the life sciences, BEN's core content aims to jump-start teaching introductory biology courses by unifying resources that are otherwise highly fragmented and widely dispersed.

The Pathway builds on BEN's successful track record as a portal manager providing database development, resource cataloging, metadata validation software tools, and Web trend reporting for professional societies. BEN's Learning Object Management (LOM) cataloging system has seven components:

- General
- Lifecycle
- Technical
- Educational
- Rights
- Classifications (subject taxonomy and pedagogic use taxonomy)
- Metadata

---

[135] Information about the Biological Sciences Pathway is complied from documents from the new collaborator's meeting, February 1-3, 2006 available at BEN:
http://www.biosciednet.org/project_site/ben_collaborator_meeting_feb06.php.

In addition to developing digital libraries with common technical standards that contribute resources to the BEN portal, BEN partners promote best practices for pedagogy, authentic assessment and the development of multidisciplinary biological sciences resources. A shared online workspace facilitates communication among collaborators. BEN relies on NSDL's technical architecture for integration of its resources into the NSDL Data Repository as well as access to NSDL's new applications (e.g., Expert Voices, Content Alignment Tool).

Table 21: BEN Partner Libraries[136]

Existing Digital Libraries	New Digital Libraries
• AccessExcellence.org: National Health Museum http://www.accessexcellence.org/ • APSArchives.org: American Physiological Society http://www.apsarchive.org/Main/index.asp • BioMoleculesAlive.org: American Society of Biochemistry and Molecular Biology http://www.biomoleculesalive.org/ • EcoEd.net: Ecological Society of America http://www.ecoed.net/ • MicrobeLibrary.org: American Society of Microbiology http://www.microbelibrary.org/ • Science's STKE: American Association for the Advancement of Science http://stke.sciencemag.org/	• AIBS: American Institute for Biological Sciences http://www.actionbioscience.org/ • BCC: BioQuest Curriculum Consortium http://www.bioquest.org/ • BSA: Botanical Society of America http://www.botany.org/ • DNALC: Dolan DNA Learning Center http://www.dnalc.org/ • EntDL: Entomology Digital Library http://cipm.ncsu.edu/PIinfo.cfm?PIID=10062003024214 (under development) • SDB: Society for Developmental Biology http://www.sdbonline.org/ • VIDA: Video and Image Data Access (VIDA)/Cal State Fullerton http://scied.fullerton.edu/vida/vidapedagogy.html

To ensure quality control of learning object resources, BEN partner societies are expected to establish a peer review framework that specifies the review timeline, criteria, ranking, and types of reviewers involved in evaluating each type of resource. Examples of the peer-review processes created by its constituent professional societies are available from BEN's Web site.[137] While the number of BEN resources is relatively low at present, it is one of the few NSDL projects with a coherent cohort of peer-reviewed individually tagged lesson plans and classroom activities. As January 2006, BEN's inventory of 4,111 resources included:

---

[136] BEN Partners list with links to services available from http://www.biosciednet.org/portal/about/partners.php.

[137] Peer Review Process of BEN Partners available from http://www.biosciednet.org/project_site/PeerReviewProcessOfBENPartners.pdf.

- AAAS (220 lesson plans and multimedia resources)
- ABLE (66 Lab Exercises and Manuals; 2 Teaching Strategies)
- AIBS (184 teaching and learning resources)
- APS (501 teaching and learning resources)
- APSNet (57 Plant Disease Lessons and articles)
- ASBMB (39 articles and interactive resources)
- ASM (1141 teaching and learning resources)
- BSA (948 annotated images)
- ESA (192 teaching and learning resources)
- FUN (20 journal articles)
- HAPS (266 journal and newsletter articles)
- NHM-Access Excellence (206 teaching and learning resources)
- STKE (317 reviews, perspectives, and multimedia resources)
- SOT (9 teaching and learning resources)[138]

BEN's user interface supports basic keyword and advanced searches along with browsing by subject and resource type. The number of items represented in each of the 76 subject areas ranges from microbiology and botany with more than 1,000 resources to hematology and glycobiology with fewer than five. The 44 categories of resource types span from images (1,352 items) and journal articles (747 items) to maps, discussion groups and assessment-exam with answer key (1 item). Advanced search offers a variety of filters, described in the previous report. As the aggregation of cataloged resources grows, the utility of these filters will increase. BEN's User Survey, conducted in September 2004, found that users (550 responses) accessed all the BEN partner sites almost equally; 56 percent downloaded resources and 67 percent used BEN resources for lectures.[139]

In four years time, BEN expects to have established 45 college and university faculty representatives around the country who are trained to provide assistance to prospective BEN contributors and users. BEN operates under the aegis of a Coordinating Council that includes representatives from the AAAS and four professional societies as well as a national Advisory Board comprised of college and university educators.

---

[138] Information from Yolanda George's presentation at the new collaborator's meeting, February 1-3, 2006:
http://www.biosciednet.org/docs/BEN_New_Collaborators_Feb06/George%20BEN%20February%202006.pdf.

[139] BEN User Study, (Chang et al.) September 2004, available from http://www.biosciednet.org/project_site/BEN_Survey_Article_October_2004.pdf.

4.3.4 DLESE: Digital Library for Earth System Education

Update Table 17: DLESE based on DLF Survey responses, Fall 2005

	DLESE: Digital Library for Earth System Education http://www.dlese.org/
ORGANIZATIONAL MODEL	Community-based organization with NSF funding.
SUBJECT	Science: Geosciences
FUNCTION	Information system and services to facilitate learning about the Earth system at all educational levels.
PRIMARY AUDIENCE	Educators
STATUS	Established
SIZE	12,000 learning resources in > 20 collections, continually growing. Includes community-contributed teaching tips, resource reviews, and news and opportunities announcements.
USE	Per month: 50,00 user sessions
ACCOMPLISHMENTS	1. Ongoing accessioning of multiple collections. 2. Services-oriented architecture (SOA) including Web search service and java script search that allows for customized search interfaces and greater dissemination of resources (Weatherley 2005). 3. Distributed, Web-based cataloging tool that supports multiple collections and multiple metadata frameworks. 4. OAI data provider and harvester tool.
CHALLENGES	1. Strategic planning 2. Continuing to meet the emerging needs of the geosciences education community. 3. Connecting with other geoscience cyberinfrastructure initiatives that will help integrate research and education.
TOOLS OR RESOURCES NEEDED	No response.
GOALS OF NEXT GENERATION RESOURCE	See above.

Funded by NSF's Directorate for Geosciences, the DLESE Program Center (DPC) operates under the aegis of the University Corporation for Atmospheric Research (UCAR) in Boulder, Colorado. DLESE plays a leadership role in bridging the education and research components of geoscience cyberinfrastructure (Marlino et al 2004).

The goals of the DLESE Program Center are to:

- *develop and provide library infrastructure tailored to specific geoscience education needs;*
- *enable distributed collections and services to act as an integrated whole;*
- *provide interoperability services with other library efforts (e.g. NSDL);*

- *support community capacity building by providing tools, components, and services that enable the development of high-quality collections of teaching and learning resources;*
- *conduct ongoing library operations; and*
- *offer broad-based community support.*[140]

DLESE serves both K-12 science instruction and undergraduate education. According to a short user survey conducted from October 2004 through February 2005, 34 percent are K-12 science teachers and 12 percent college/university faculty members; 13 percent are K-12 students and 10 percent are college students. Developers of educational materials account for 7 percent; parents, librarians, and others (non-geoscience teachers, outreach coordinators, professional development experts, and DLESE staff) comprise the remaining 24 percent.[141]

What are they seeking?
- 30 percent   Materials for students
- 18 percent   Materials for an assignment
- 13 percent   Information about the library (i.e. DLESE)
- 7 percent   Information for curriculum development
- 6 percent   Information for their own learning
- 5 percent   Collaborators for a project[142]

DLESE maintains two primary collections. Resources in the "DLESE Community Collection" (~7,100 items) meet basic guidelines in terms of subject relevance and functionality.[143] The more selective "DLESE Reviewed Collection"[144] is composed of resources (~670 items) that have been evaluated against seven criteria:

1. scientific accuracy;
2. pedagogical effectiveness;
3. completeness of documentation;
4. ease of use for teachers and learners;
5. ability to inspire or motivate learners;
6. importance or significance of the content, and
7. robustness as a digital resource. (Kastens et al. 2005)

---

[140] As articulated in "About DLESE: Overview," available from http://www.dlese.org/about/.
[141] http://www.dlese.org/cms/evalservices/recent_developments/recent provides an overview of the final set of data. Number of respondents =524
[142] From Overview report available from:
http://cybele.colorado.edu/docs/DLESE_User_Survey_Overview_1_19_04.pdf
[143] DLESE Community Collection Scope Statement is available from
http://www.dlese.org/Metadata/collections/scopes/dcc-scope.htm.
[144] DLESE Reviewed Collection Scope Statement is available from
http://www.dlese.org/documents/policy/CollectionsScope_final.html.

In addition, DLESE collects metadata from other digital libraries (e.g., Alexandria Digital Library) and thematic collection developers (e.g., the Digital Water Education Library—DWEL or the Earth Exploration Toolbook—EET). The EET is an innovative collaboration that utilizes earth science data within NSDL and DLESE to create an online collection of computer-based learning problem-solving activities.[145] Currently EET has fourteen chapters organized around learning activities, such as analyzing the Antarctic Ozone Hole, exploring regional differences in climate change, or visualizing carbon pathways. Each chapter is accompanied by relevant datasets (derived from NASA, USGS, and U.S. Census data or other sources) and technology tools (e.g., GIS, image processing programs, spreadsheet applications). To facilitate use, EET has created companion, professional development Data Analysis Workshops for teachers.

Users can browse DLESE's collections by subject, resource type, and grade level.

Figure 28: Screenshot of DLESE Reviewed Collection by Grade Level
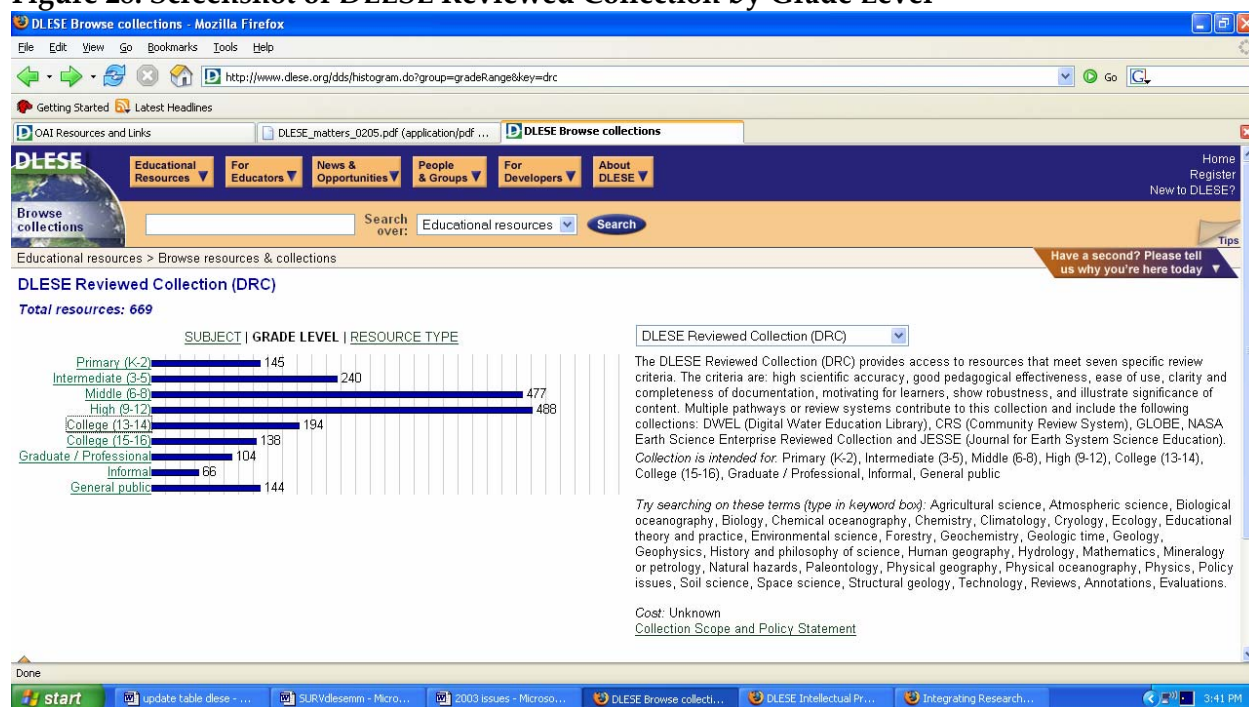
Source: http://www.dlese.org/dds/histogram.do?group=gradeRange&key=drc (May 2006)

As illustrated in the figure above, each collection has a bar graph, charting the number of resources as well as a collection annotation and link to the collection's scope and policy statement.

---

[145]Earth Exploration Toolkit proposal abstract available from http://nsdl.org/resources_for_library_builders/projects.php?pager=projects&this_sort=start_date&keyword=&project_id=0532881.

Text searches can be filtered by grade level, resource type, collection, and educational standard. At present, DLESE has the ability to search by National Science Education Standards (NSES) and by National Geography Standards (NGS). The National Geography Standards organize learning concepts under six broad topical categories: Environment and society, Human systems, Physical systems, Places and regions, the Uses of geography, and the World in spatial terms, for a total of 18 individual standards. The NSES are hierarchical and permit users to choose grade level, broad topic, and learning goal. For example:

- Grades 9-12
    - Earth and space science
        - Energy in the earth system
        - Geochemical cycles

DLESE is working jointly with Syracuse University's Center for Natural Language Processing (CNLP) to incorporate additional state and national standards into the library and connect with the Achievement Standards Network (ASN) database maintained by JES & Co. and funded by the NSDL. In addition, the CNLP released a prototype of its Content Assignment Tool (CAT) as an API integrated within DLESE's Collection System (DCS) in early 2006 (Diekema and Devaul 2006). CAT uses natural language processing to analyze the content of learning resources, such as lesson plans, and then automatically suggests relevant national and/or state standards. It is intended not only to aid catalogers in assigning appropriate standards and providing a cross-walk between different state and national standards, but also permits users to save their choices to a database. A beta version is currently available for testing by registered users.[146]

DLESE makes innovative use of its "Community Review System" to create customized reports for teachers that assess the effectiveness of digital learning resources in their classrooms (Kastens and Holzman 2006). *The Introductory Geoscience Virtual Textbook* was created as a test bed for the CRS individualized teacher report system, utilizing DLESE resources to teach students about basic concepts in Earth science.[147] Both students and the instructor write reviews of the digital resources based on the seven criteria noted above for "reviewed resources" and then the DLESE CRS creates a report aggregating and comparing the data from the instructor's and student's perspective. Examples of various types of reports generated by the CRS are available at DLESE's Web site.[148]

---

[146] Information available from Press Release of December 9, 2005:
http://www.eotepic.org/modules.php?op=modload&name=News&file=article&sid=584.
[147] *The Introductory Geoscience Virtual Textbook* by Christopher DiLeonardo (September 2004) is available from http://crs.dlese.org/testbed/Textbook/index.html.
[148] See the Community Review System annotations and reports available from http://crs.dlese.org/annotations/.

Figure 29: Screenshot of DLESE search for learning resources about "hurricane" that meet National Science Education Standards.
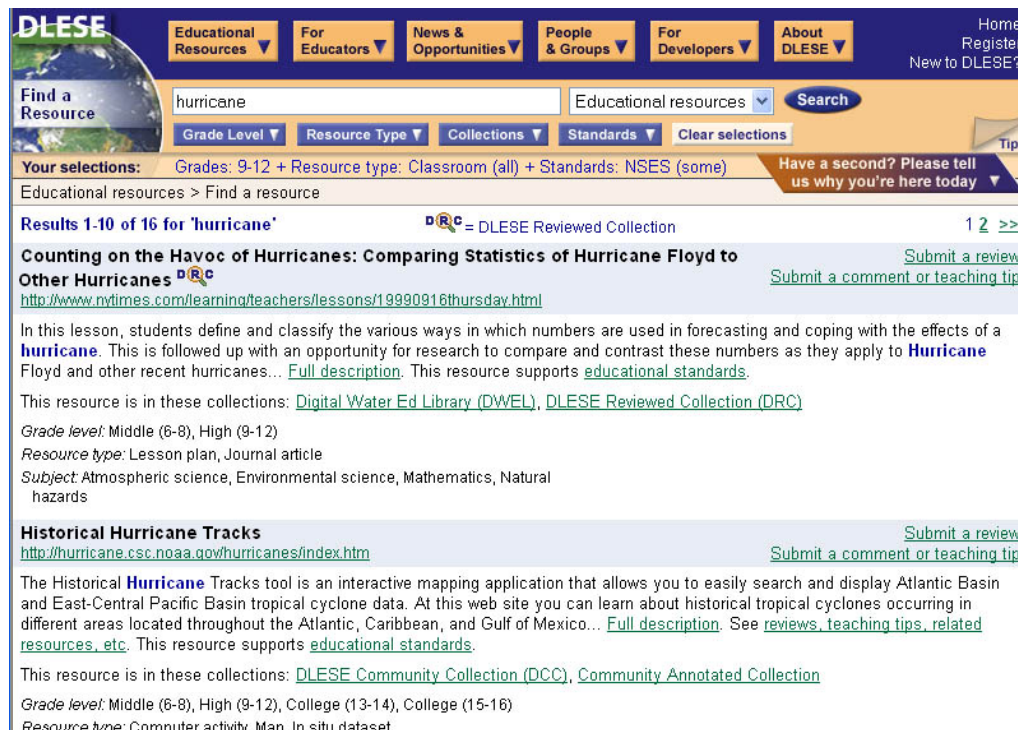
Source: http://www.dlese.org/ (May 2006)

Since the 2003 DLF report was issued, DLESE's information technology infrastructure has evolved into a service-oriented architecture (SOA), with improved interoperability capabilities that extend its reach through Web service and JavaScript APIs (Weatherley 2005) (see http://www.dlese.org/dds/services/). The Center for Ocean Science Education Excellence (COSEE, http://www.cosee.net/), for example, has embedded a custom DLESE search in their Web portal that is implemented using the DLESE Search Web Service and the My NASA Data portal utilizes a custom search page implemented with the JavaScript API (http://mynasadata.larc.nasa.gov/DLESE_search.html). In addition to these, DLESE services and APIs are being used to deliver DLESE resources interactively to users of GLOBE, NASA S'COOL, the GEON portal and several other institutional Web sites and portals.

The California Digital Library is harvesting DLESE's OAI records and integrating them into a geosciences portal tailored to the users of the UC campus libraries. DLESE is also a Principal Investigator (PI) Institution in GEON, a network building cyberinfrastructure capacity in geoinformatics for research (GEON) and educational (DLESE) purposes.

> *GEON is based on a service-oriented architecture (SOA) with support for "intelligent" search, semantic data integration, visualization of 4D scientific datasets, and access to high performance computing platforms for data analysis and model execution -- via the GEON Portal.* http://www.geongrid.org/

GEON and DLESE interoperate in a number of important ways (Wright 2004). GEON uses the ADN Metadata Framework, (jointly developed by the Alexandria Digital Library, the NASA Science Mission Directorate and DLESE)[149] and the two services share collection records. GEON Web services and content are available in DLESE (http://geon01.dlese.org/) and the GEON Portal provides access to DLESE. GEON, DLESE, and the University of Colorado are collaborating to create an "Educational Knowledge Organization System" (EKOS) that supports conceptual browsing (concept strand maps) to align learning outcomes and educational standards with DLESE's resources (Wright 2004, Sumner et al. 2004).[150]

Figure 30: Screenshot Preview of "Browsing Earth Concepts"

Source: http://preview.dlese.org/jsp/cms/ (May 2, 2006)

Custard and Sumner (2005) report on their research to "Using Machine Learning to Support Quality Judgments" about digital resources and collections. NSDL and DLESE were used as a test case for their research to determine if a set of "indicators could be used to accurately classify resources into different quality bands and to determine which indicators positively or negatively influenced resource classification." According to the authors, "The results suggest that resources can be automatically classified into quality bands, and that focusing on a subset of the identified indicators can increase classification accuracy." In the future, collection curators may rely on these "next

---

[149] An overview of the ADN Metadata Framework is available from http://www.dlese.org/Metadata/adn-item/history.htm.

[150] For more information refer to http://geon01.dlese.org/projects/strandmap.html.

generation cognitive tools" to support their qualitative decisions about which digital resources to acquire.

Publications and presentations by members of the DLESE community are listed in the bibliography maintained at the DLESE Web site.[151]

4.3.5 MERLOT: Multimedia Educational Resource for Learning and Online Teaching

Update Table 18: MERLOT based on DLF Survey responses, Fall 2005

	MERLOT: Multimedia Educational Resource for Learning and Online Teaching http://www.merlot.org/
ORGANIZATIONAL MODEL	Community-based with free open individual or partner membership (with annual institutional fee-based benefits).
SUBJECT	Multi-disciplinary
FUNCTION	Improve the effectiveness of teaching and learning by increasing the quantity and quality of peer reviewed online learning materials that can be easily incorporated into faculty-designed courses.
PRIMARY AUDIENCE	Academic community
STATUS	Established
SIZE	2 Sustaining institutions; 23 system and campus partners and affiliates; 13 professional societies and 9 digital libraries; 8 corporate sponsors and 30,000 individual members. 13,000 learning materials (37% growth) organized in 15 disciplines categories.
USE	Daily use with 1,000 new members monthly.[152]
ACCOMPLISHMENTS	1. Established reputation for high quality and sustainability. 2. Development of Corporate Partnerships. 3. Development of JOLT (Journal of Online Learning and Teaching). 4. Provision of discipline-communities.
CHALLENGES	1. High demand but limited resources.
TOOLS OR RESOURCES NEEDED	Web browser

---

[151] The DLESE bibliography is available from http://www.dlese.org/documents/bibliographies/DLESE_bibliography.html.
[152] According to the CSHE study, "MERLOT does not normally report usage statistics by month. From January 1 to November 30, 2005, MERLOT reported a total of 30,232 registered users and 758,754 visits (an average of 2,273 visits per day, with an average of 10 pages per visit)" (Harley et al. 2006, 144).

GOALS OF NEXT GENERATION RESOURCE	1. Increase membership and collection growth. 2. Expansion of faculty development services. 3. Extending the disciplinary model to additional areas of academic and workforce interest.

Those new to MERLOT have several options to familiarize themselves with its services and features. From MERLOT's Web site, users can access a brief video introduction (replete with faculty testimonials), listen to a presentation about MERLOT co-sponsored by the TLT Group, watch an interview with MERLOT's Executive Director, Gerry Hanley, or listen to his longer video presentation, "Sharing Learning Objects: Serving MERLOT to Higher Education."[153] In summarizing what makes MERLOT work effectively, Hanley emphasizes these characteristics:

- We create a common means to individual ends.
- You get more than you give.
- You have a fair share in decision-making and participation.
- We hold true to academic values.
- We provide visibility, accountability and sustainability.
- You trust us to deliver high quality services.

MERLOT has an organizational partnership structure that defines levels of participation and obligations, including annual membership fees and in-kind support.[154] There are three broad organizational categories:

- higher education institutions
- non-profit institutions (professional societies and digital libraries)
- corporations

And four levels of participation:

- Affiliate: joint advocacy but low-level of cooperation, requires an application or MOU
- Project-level: collaborate in MERLOT initiatives and pay $6,500 annual fee (for campuses or negotiated rate for other organizational types) with in-kind support required for projects.
- Community: participate in MERLOT leadership and collaborate on projects; pay a $25,000 with $50,000 to $100,000 in-kind support required for leadership and initiatives.

---

[153] MERLOT's Tasting Room provides access to the video presentations. Available from http://taste.merlot.org/.
[154] An Overview of MERLOT's Partnerships is available from http://taste.merlot.org/participating/partner/partner_sum05-06.pdf.

- Sustaining: lead a MERLOT initiative, participate in MERLOT management and pay a $50,000 annual fee plus $250,000 in-kind support required.

The Partnership Comparison Chart provides details of membership benefits for institutions of higher education in the areas of training, involvement in MERLOT leadership, collaboration and evaluation opportunities, and access to MERLOT member-only resources.[155]

Since 2003, MERLOT has expanded its international outreach and content through strategic alliances in Canada, Europe and Australia. CLOE, the Co-operative Learning Object Exchange led by the University of Waterloo (Ontario, Canada) is now a major sustaining partner alongside the California State University. The ARIADNE Foundation for the European Knowledge Pool, a distributed network of learning repositories, has become a MERLOT partner. In addition, MERLOT, ADRIADNE and EdNA, the Education Network Australia of learning repositories, each offer federated searches across their collections individually or collectively.[156] They are also all members of the consortium, GLOBE (Global Learning Objects Brokered Exchange), along with eduSourceCanada and the National Institute of Multimedia Education (NIME) in Japan.[157]

MERLOT has also strengthened its corporate partnerships, which include O'Reilly Media and Sun Microsystems, three learning management systems (ANGEL Learning, Blackboard/WebCT, Desire2Learn) and two library systems (Ex Libris Ltd. and Sentient Learning).[158] These partnerships result in mutually beneficial services such as the seamless integration of MERLOT resources via Blackboard and ANGEL.[159] A similar service with WebCT will be available in July, 2006. As a matter of principle, MERLOT only signs non-exclusive agreements with vendors. It has assigned different values to its functions as follows:

---

[155] MERLOT's Partnership Comparison Chart is available from
http://taste.merlot.org/participating/partner/partner_compare05-06.pdf.
[156] Federated Search pages available from Globe
http://globe.edna.edu.au/globe/go/pid/2
MERLOT http://fedsearch.merlot.org/search.jsp
EdNA http://www.edna.edu.au/edna/search?SearchMode=Advancemode
Ariadne requires user registration and log-in (free) to search
http://ariadne.cs.kuleuven.be/silo2006/Welcome.do.
[157] Information about GLOBE is available from http://globe.edna.edu.au/globe/go/pid/2
[158] Information about MERLOT's collaborations with Learning Management Systems is available from http://taste.merlot.org/initiatives/lms.htm.
[159] For more information about the ability to search MERLOT within ANGEL refer to http://www.angellearning.com/products/lms/lor/.

- Basic Search – gratis
- Basic RSS feeds – gratis
- Advanced Search - nominal fee
- Customized RSS feeds - negotiated fees
- Federated Search - negotiated fees
- Other Services - negotiated fees

Participating vendors are required to adhere to the standard MERLOT Metadata Services Agreement in which MERLOT maintains control over the use of its technology, preventing institutions from harvesting its metadata. MERLOT's metadata is IEEE LOM or IMS metadata compliant, but it is not OAI-compliant nor is it available for export. As a result, MERLOT is only represented in NSDL at the collection-level. MERLOT does offer a search service, which is a Web service that allows remote searching of the MERLOT metadata and returns results in an XML format for display by the requester (as in the case of MERLOT's initiative with Blackboard).

In July 2005, MERLOT inaugurated *JOLT: Journal of Online Learning and Teaching* as a peer-review, open access vehicle to promote the scholarship of technology-enabled teaching and learning in higher education. *JOLT* serves as another forum in which the MERLOT community can express and examine issues of common concern.

MERLOT offers various avenues for users to keep abreast of recent developments besides its "What's new" page, quarterly email newsletter the *Grapevine*, and press releases. It supports syndication (RSS), and from the home page, users can quickly link to the most recently added resources (225 items), new member profiles (845), and peer-reviewed resources (26) contributed in the last thirty days.

MERLOT's basic user interface is the same as reported in 2003, but there a number of new or previously unrecorded features. The Advanced search functions permit users to limit their queries by a number of unique qualifiers going well beyond subject, material type, technical format, language, audience and cost. These include: learning management system compatibility (Blackboard/WebCT, Desire2Learn), iPod items, Section 508 compliant items (conform to minimal disability access standards), copyright restrictions, and availability of source code. In addition searches can be restricted to peer-reviewed resources (further refined by minimum rankings), member comments (further refined by user rankings), availability of assignments, and author snapshots. Author snapshots utilize the KEEP Toolkit developed by the Carnegie Foundation for the Advancement of Teaching to produce an illustrated synopsis (e-portfolio) of the educator's rationale, motivation, and impact on teaching and learning in developing the resource.

It is worth noting that some of these filters restrict the results to a very limited sub-set. For example, whereas 85 to 90 percent of the resources have been reviewed by faculty,

only about 15 percent actually have published "peer reviews" in MERLOT (<2,000 items).[160] According to MERLOT representatives the comparatively low proportion of peer-reviewed resources is attributable to a combination of factors including the amount of time required by faculty, the author's consent, and the quality of the material. Consequently, resources deemed of lesser interest do not receive MERLOT Peer Review. Results can be sorted by five different variables (title, author, date entered, rating, item type). It is possible to conduct sub-searches within the result set.

Figure 31: Screenshot of MERLOT's Federated Search page
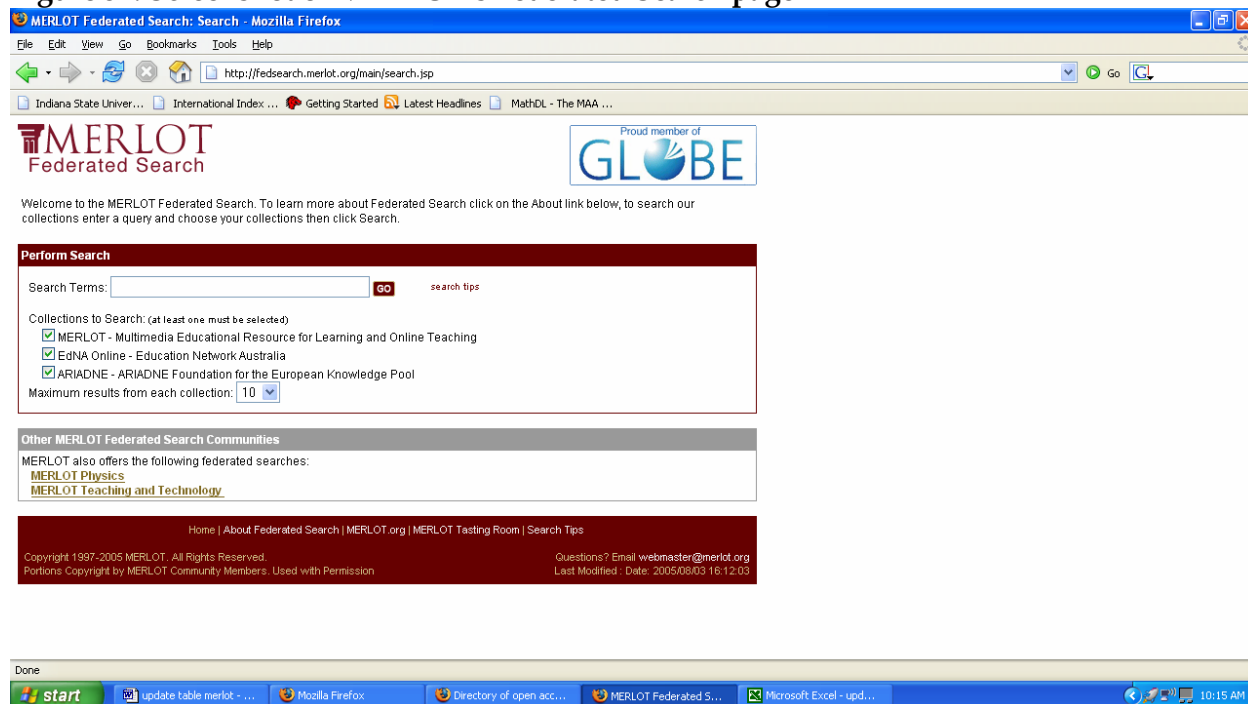
Source: http://fedsearch.merlot.org/main/search.jsp (March 2006)

In addition to federated searches across ARIADNE and EdNA's learning object repositories, MERLOT offers two subject-based federated searches: physics (covering MERLOT Physics and ComPADRE—Digital Resource Collections for Physics and Astronomy Education developed as a NSDL Pathway) and teaching and technology (covering MERLOT resources and the University of Carolina's Professional Development Portal). Currently in test is a federated search from the MERLOT Information Technology portal into IEEE Computer Society's extensive digital library (http://www.computer.org/). A new version of the MERLOT Web site is currently under development and planned for release at the MERLOT International Conference in August, 2006.

---

[160] Email correspondence of April 4, 2006 from Martin J. Koning Bastiaan, Director of Technology, MERLOT.

4.3.6 Current Issues and Future Directions

Each in their own way, these services face organizational challenges to increase content and usage. NSDL is developing "pathways"—exemplified by NEEDS and BEN—to focus resources for particular audiences and coalesce services across sectors. NEEDS is merging with TeachEngineering to serve the full spectrum of K-12 to lifelong learners. BEN has excelled at developing models for transforming smaller organizations to become contributors to digital libraries. DLESE is developing tools to support distributed cataloging of multiple collections and different metadata frameworks. MERLOT is bringing in new international and corporate partners. This cohort has developed a number of effective marketing and outreach vehicles to secure and extend their user base:

> NSDL now offers more interactive communications features from its Web site and is organizing more teacher workshops;
> NEEDS offers digital repository services to other organizations; oversees the Premier Award to recognize outstanding courseware, and displays monthly theme pages;
> BEN is developing a faculty campus representative program;
> DLESE's supports a distributed, Web-based cataloging tool and is working with NSDL and Syracuse University to incorporate state and national standards into its database; and
> MERLOT inaugurated *JOLT: Journal of Online Learning and Teaching* as a peer-review, open access vehicle to promote the scholarship of technology-enabled teaching and learning in higher education.

An essential ingredient to their success is offering quality assurance of content, one aspect of which is peer review. The user interfaces of SMETE/NEEDS, DLESE and MERLOT support filters to peer-reviewed items. However the actual proportion of such items is relatively low. A search limited to peer-review resources returns only 25 results in SMETE or NEEDS. Less than 15 percent of MERLOT's are peer-reviewed, whereas only 700 of DLESE's 12,000 resources are part of its reviewed collection. Although BEN has made considerable gains (310 percent increase since 2003), peer-reviewed resources constitute less than 15 percent of its database as well. This suggests that peer-review in the digital realm is still at an early stage of acceptance and is not well-integrated into faculty traditions and reward systems. According to an NSDL study underway by Alan Wolf, "The science faculty that he studies claim to trust neither peer review nor community vetting; instead, they simply rely on their own personal judgment in every

case of using an OER [online educational resource], or they consult with a trusted colleague" (Harley et al. 2006, 166)[161].

These services also face the challenge of meeting the diverse needs of an expanded user base, particularly those that attempt to span the K to grey clientele. Research studies sponsored by NSDL among others reveal considerable differences by education sector in terms of what teachers need to integrate digital resources into their pedagogy (California Digital Library 2004, Hanson and Carlson 2005, Harley et al. 2006). NSDL's pathways are intended to target resources and services to particular audiences, but it remains to be seen if these services can effectively serve diverse and sizeable constituents which have widely varying needs and operate in different conditions. NSDL, in particular, notes the "great diversity in evaluation methods and tools across 190+ NSDL digital library projects." This is corroborated by the CSHE study which reports that six NSDL collections included in their review "used almost completely different metrics to describe themselves and their use" (Harley et al. 2006, 157).

While these services are making strides to integrate their resources into other services (e.g., NSDL's incorporation into academic library portals and science.gov; MERLOT's federated search system and partnerships with WebCT/Blackboard; DLESE's partnership with GEONgrid), it remains to be seen how they will join up with other national and international communities of practice formed around e-learning technology platforms and e-learning frameworks. How do their efforts mesh, for example, with international efforts to make content object repositories interoperable such as CORDRA (Content Object Repository Discovery and Registration/Resolution Architecture, http://cordra.net/) or the IMS Global Learning Consortium (http://www.imsglobal.org/) (Kraan and Mason 2005)?

Finally, financial sustainability is a major challenge, cited particularly by NSDL and MERLOT, but also evident in responses from the other services. Through the efforts of its Sustainability Standing Committee, NSDL is tackling this issue by formulating a decision-tree and providing its constituent projects with information about establishing marketing and business plans; however, NSDL as a whole—like other services in this report—attest to the need for more public and private funding options. The California Digital Library's market assessment of NSDL suggests that "academic libraries see limited value in another Web science portal, but would be willing to consider paying for deep integration with their existing search tools" (CDL 2004, 3). Even MERLOT, which has a fee-based membership structure, identifies the challenge of "high demand, but limited resources." Nor can MERLOT count on maintaining its current membership base. The CSHE study of "Use and Users of Digital Resources" notes that while MERLOT (alongside a handful of other services) "could function on an existing base of

---

[161] As mentioned elsewhere in this earlier in this report, Alan Wolf and Flora McMartin are using CSHE's research design to study comparable issues of use in STEM disciplines. See NSF project ID 435398, awarded January 1, 2005, "Faculty Participation in NSDL—Lowering the Barriers."

support, budgetary volatility encouraged them to continuously watch for new funding opportunities" (Harley et al. 2006, 147).

4.4   Joining Forces: Cultural Heritage and Humanities Scholarship

At present, we have the opportunity to reintegrate the cultural record, connecting its disparate parts and making the resulting whole available to one and all, over the network. . . . Like most grand challenges, this one can be simply stated: make it possible for people to explore the totality of our accumulated global cultural heritage, now scattered throughout libraries, archives, or museums.   ACLS, Cyberinfrastructure in the Humanities & Social Sciences, 2005

The eleven services under review in this section serve as exemplars of ways in which librarians, archivists, educators, and scholars are collaborating to build digital collections and tools in support of cultural heritage and humanities scholarship. The discussion begins with two services that bridge the cultural divide by presenting collections and content from libraries, museums, and archives in a unified way. Cornucopia, sponsored by the Museums, Archives & Libraries Council (UK), serves as a single point of access for resource discovery, based on 6,000 collection-level descriptions from 2,000 institutions in the UK. Since the 2003 DLF report appeared, Cornucopia began to make its collection metadata available via OAI and SOAP. Further, it served as a model for a new project in the US led by the University of Illinois, namely the IMLS Collections and Content gateway to digital projects funded by the IMLS National Leadership Grant Program. The Institute of Museum and Library Services (IMLS) is an independent grant-making agency of the federal government whose mission is "to lead the effort to create and sustain a 'nation of learners'" (http://www.imls.gov/).

Both projects use the RSLP (Research Support Libraries Programme) Collection Level Description (CLD) Metadata Schema which enables consistently formatted descriptions to be created and linked through parent-child relations and association relationships (as depicted in Figure 32), building on entity relation models for collection descriptions (Healey 2000, 2005)[162]. In addition, these projects are informed by NISO's (National Information Standards Organization) "A Framework of Guidance for Building Good Digital Collections" (2nd edition, 2004) and the NISO Metadata Initiative, described in the next section of this report.[163]

---

[162] Information about the RSLP CLD Schema and an online tutorial are available from http://www.ukoln.ac.uk/cd-focus/ and http://ukoln.ac.uk/metadata/rslp/schema. Healey's studies about analytical models for collections and their catalogs are available from http://www.ukoln.ac.uk/cd-focus/model-ext/intro.html.
[163] Respectively available from http://www.niso.org/framework/Framework2.html and http://www.niso.org/committees/MS_initiative.html.

The DLF's Digital Collections Registry, which is maintained also by the University of Illinois, is briefly described before turning to three services included in the 2003 DLF survey: the Library of Congress's American Memory, the Sheet Music Consortium, and the Collaborative Digitization Program's (formerly Colorado Digitization Program) Heritage West (formerly Heritage Colorado). These represent various models of fostering cooperative digital collections and aggregating at the international, national, and regional level.

Two pilot projects—The American West and DLF Aquifer—sponsored by the California Digital Library and the Digital Library Federation respectively, are starting to put into practice many of the lessons learned from previous collaborative projects. They are pooling digital content and building tools and services targeted to particular audiences. Meanwhile, Emory University's capstone initiative, SouthComb, leverages its prior digital initiatives including AmericanSouth covered in the 2003 DLF survey, to create a scholarly portal for Southern Studies.

Two scholar-driven projects round out this section. Since 2003, the Perseus Digital Library (PDL) has rebuilt its text system, released a new Web site, and launched a named entity browser. It plans to migrate its core data to the Tufts Institutional Repository in order to concentrate on research and development activities. Once PDL research applications prove viable, they will move to the IR's production server. Finally, NINES (Networked Interface for Nineteenth-Century Electronic Scholarship) represents a new scholar-driven model of aggregating peer-reviewed work and presenting it for use along with a suite of interpretative digital tools. Led by Jerome McGann, the John Stewart Bryan University Professor at the University of Virginia and editor of the acclaimed Rossetti Archive, NINES has garnered endorsements from five disciplinary societies and a host of other influential humanities computing organizations and projects.

4.4.1 Cornucopia

Update Table 19: Cornucopia based on DLF Survey responses, Fall 2005

	Cornucopia http://www.cornucopia.org.uk/
ORGANIZATIONAL MODEL	Museums, Archives & Libraries (MLA) Council (UK)
SUBJECT	Cultural heritage
FUNCTION	A single point of access for resource discovery based on collection level descriptions.
PRIMARY AUDIENCE	General public
STATUS	Established
SIZE	> 6,000 collection descriptions from 2,000 institutions.
USE	Not available
ACCOMPLISHMENTS	1. Growth of contributions and descriptions. 2. Facility of locations to create/maintain their own descriptions. 3. Availability of data via OAI and SOAP.
CHALLENGES	1. Funding 2. Reconciliation of RSLP CLD Schema with different sector schemas. 3. Standardization of terminology
TOOLS OR RESOURCES NEEDED	No response
GOALS OF NEXT GENERATION RESOURCE	1. Integration with Archive collections. 2. Extension of SOAP target

Developed by the Museums, Libraries and Archives Council (MLA), Cornucopia is a searchable database of some 6,000 collection descriptions emanating from 2,000 cultural heritage institutions in the UK. In spring 2004 Cornucopia migrated to a new software system and realigned almost all of its data structure to conform to the RSLP (Research Support Libraries Programme) Collection Level Description Metadata Schema (Turner 2005). The new system enhanced Cornucopia's functionality. Contributors can now edit and enter their collection data through a Web-based direct entry client; moreover, Cornucopia's data became available for OAI harvesting and Web service access. This enables interoperability among cultural heritage sites in the UK. For example, the People's Network Discover Service (http://www.peoplesnetwork.gov.uk/discover/) is harvesting Cornucopia data and making it searchable as one component of an aggregation harvested from an increasing number of cultural heritage sources. MLA's longer term vision is to provide integrated access to a wide range of data from the cultural sector, in which Cornucopia figures prominently.

As Cornucopia expands to incorporate more heterogeneous resources from an expanded institutional (e.g., including many more library collections) and user base, UKOLN

undertook a strategic review of indexing options. A series of reports issued in September 2005 and January 2006 present comparative analyses of alternative thesauri, name authority files, and controlled vocabularies; recommend preferred indexing conventions for Cornucopia; and outline action plans for implementation. Among the key recommendations are to use the UK Archival Thesaurus (UKAT) for subject indexing and to abandon Cornucopia's current place indexing and use certain sections of the UNESCO Thesaurus instead. The time browsing page will be overhauled and new audience values and collection strength information added. New "Contributor Guidelines" give examples of how to assign appropriate index terms for subjects, places, time periods, names, audience levels, and collection strength based on UKOLN's findings.[164]

Cornucopia's search and retrieval features have improved since 2003; however, in view of the new indexing recommendations, the description of its current functionality is provisional. Collections can be browsed by seven categories: time, people, place, subject, culture (e.g., Ancient Greece, Jewish, Maya, Viking), and institution. The user interface supports hierarchical, faceted browsing by subject. There are 21 broad subject categories (e.g., Education, Events, Information and Communication). In advanced search mode, users can narrow a collection title search by time period, place, type of institution (library, archive, or museum), or county. Alternative keywords are suggested to expand the search, based on the UK Archival Thesaurus. Results are returned with brief annotations, and link to full records that include (a) a collection summary, (b) location details (directory information about the institution), and (c) additional collection information; in some instances, there are links to the item via the institution's catalog. The "collect me" feature allows users to gather and save search results during a session for printing or emailing.

In addition, users can perform a search by postal code to locate collections in a particular location or conduct a search within or across three other Web services, including Cecilia: Find Music Collections in the UK and Ireland; Darwin Country; and Google. At present, no explanations are given to users about the coverage of the other services. However, Cecilia is a database of some 1,800 collection descriptions of music resources held in 600 libraries, archives and museums in the UK and Ireland (http://www.cecilia-uk.org/). Darwin Country, a partnership of several regional museums, focuses on the history of science, technology and culture in the West Midlands during the 18th and 19th centuries; it is affiliated with the UK's "Curriculum Online" initiative.[165] Among other features, Darwin Country enables the exploration of artifacts consisting of nearly 12,500 historic images (http://www.darwincountry.org/).

---

[164] Series of six Cornucopia Phase 2 reports co-authored by Ann Chapman and Rosemary Russell provided to the author by Chapman via e-mail of April 25, 2006.
[165] For information about Curriculum Online refer to http://www.curriculumonline.gov.uk/.

Besides Cecilia, various other digital projects in the UK have chosen to use Cornucopia's software and will provide their own user interfaces. They include:

The Concert Programmes Project
> *This collaborative project, with Cardiff University as the lead institution, will create an online database of holdings of concert programmes held in libraries, archives and museums in the UK and Ireland, providing access to a vital source of information about musical life from the eighteenth century to the present day.*
> http://www.cph.rcm.ac.uk/Concert%20Programmes/Pages/Home%20Page.htm

Inspire
> *The ultimate aim of Inspire is to create seamless access across over 4000 public, 3 national, almost 700 higher education libraries, as well as special libraries and those in further education colleges and schools, and to build an effective interface to resources for learning with museums, galleries, archives and other organisations and services.*
> http://www.inspire.gov.uk/index.php

DiadEM
> *A regional initiative funded by the MLA and Libraries and Information East Midlands to create collection level descriptions (CLDs) of the special library collections in the East Midlands.*
> http://www.inspire.gov.uk/pdf/DiadEM%20description%20for%20INSPIRE%20website.pdf

The Egyptologists Subject Specialist Network (SSN), which will be a forerunner for several other SSNs
> http://www.mla.gov.uk/website/programmes/renaissance/Subject_Specialist_Networks/

4.4.2 IMLS Digital Collections & Content (DCC)

Update Table 20 : IMLS Digital Collections & Content based on DLF Survey responses, Fall 2005

	IMLS Digital Collections & Content http://imlsdcc.grainger.uiuc.edu/
ORGANIZATIONAL MODEL	Collaboration among UIUC Library, UIUC Graduate School of Library & Information Science, IMLS
SUBJECT	Cross-disciplinary
FUNCTION	Registry and repository with search and discovery tools for integrated access to content of IMLS National Leadership Grant (NLG) collections.
PRIMARY AUDIENCE	Academic Community
STATUS	Established
SIZE	Registry: 151 NLG collections plus 100 brief descriptions of related collections. Metadata repository: 266,000 records from 85 IMLS NLG collections.

USE	Not available
ACCOMPLISHMENTS	1. Creation of IMLS NLG Collection Registry with rich data input system & browse interface. 2. Helping NLG projects develop OAI data providers and promulgating OAI Best Practices. 3. Development of IMLS metadata repository
CHALLENGES	1. Maintenance 2. Keeping up with changing standards
TOOLS OR RESOURCES NEEDED	No response
GOALS OF NEXT GENERATION RESOURCE	1. Continue to add all new IMLS NLG grants to the Registry. 2. Continue to assist IMLS NLG grantees in setting up OAI data providers.

A collaborative initiative led by the University of Illinois, Urbana-Champaign (UIUC) Library and Graduate School of Library & Information Science, this gateway is intended to bring greater visibility and utility to digital collections funded by the IMLS (Institute of Museum and Library Services). The DCC serves as both a registry of collection-level descriptions of National Leadership Grant (NLG) projects and a metadata repository of item-level records from a subset of these collections. In its next phase of development (funded through 2007), the DCC expects to add a sample of digital collections funded via IMLS to State Library Administrative Agencies in support of the Library Services Technology Act (LSTA).

Integral to the development of the DCC, the principal investigators are testing the assumptions of the NISO/IMLS *Framework of Guidance for Building Good Digital Collections* (2004), namely how the registry and repository might serve as "infrastructure components" with "the potential to facilitate the reuse of digital content in new and different ways – by enabling more effective search and discovery across multiple collections and among and between individual information objects that will allow communities of scholarly interest to view an information landscape as best meets their needs" (Cole and Shreeves 2004, 309). Specifically, the DCC experiments with OAI-PMH interoperability best practices in terms of collection identity, metadata normalization and enrichment for specific audiences, and portal interface and functional design issues (Cole 2006).

In creating the collection registry model, the DCC draws on research about how to define and describe collections, ultimately opting to adapt the RSLP Collection Description Schema and the Dublin Core Collection Description Application Profile[166] (Cole and Shreeves 2004, 312). Taking into consideration similar projects, in which

---

[166] See footnote above about the RSLP Schema. Information about the Dublin Core Metadata Initiative is available from http://dublincore.org/groups/collections/.

Cornucopia figured prominently, the DCC arrived at a collection description metadata schema with four classes of entities:

- Collections
- NLG projects
- Institutions
- Administrators (Ibid, 317-18)

Like Cornucopia, DCC enables contributors to add and edit their collection information. The DCC Web site offers full details about the metadata schema as well as a diagram illustrating the relationship among entities. The principal investigators elaborate:

> A collection may have been created by multiple NLG projects and have multiple administrators. A collection may only have one hosting institution, but may have multiple contributing institutions. A collection may have multiple sub-collections, complementary collections, or source physical collections. A NLG project may have only one administering institution, but may have multiple participating (or collaborating) institutions.

Figure 32: Relationships Among Entities in the IMLS DCC Collection Description Metadata Schema
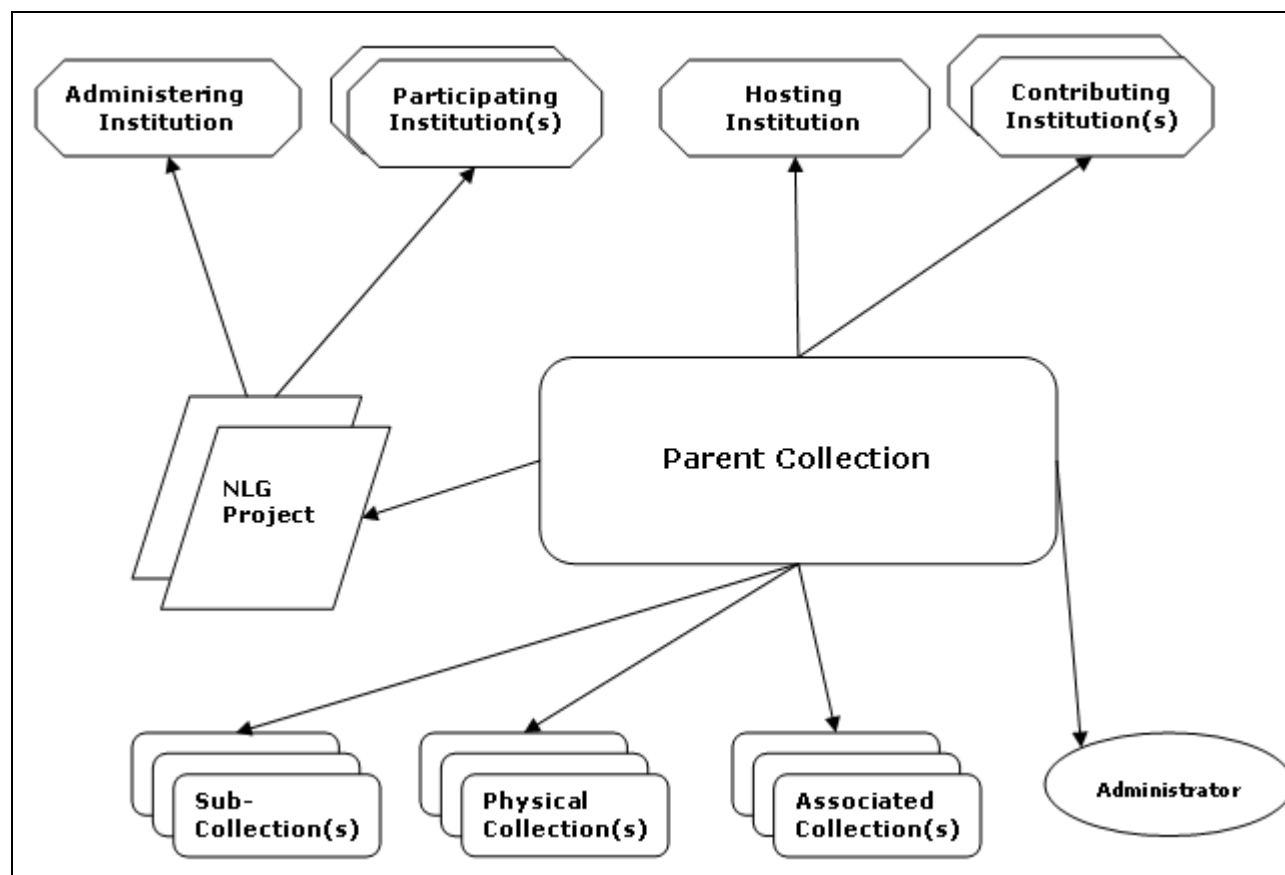
A second major component of the project involves enabling cross-collection searching of item-level metadata using OAI-PMH to facilitate interoperability. To this end, the DCC deployed several strategies to help participating collections become OAI data providers including implementing an OAI Static Repository for some projects, and working with CONTENTdm, (already in use by other projects), to support "resumption tokens" that help to control the flow of records in manageable chunks to the DCC. As result, more collections are able to contribute item-level records to the DCC. Nevertheless as of this writing, only about half of the collections have associated item-level records. Noting that the absence of item-level metadata is particularly prevalent for exhibit and learning object focused projects, Cole and Shreeves offer other reasons why NLG projects are not yet OAI-compliant:

- The digital collection is not yet public.
- The technical infrastructure is not in place.
- The technical infrastructure is in migration, for example, migrating to a new content management system.
- All collaborators in a particular project have not reached agreement to share metadata via OAI.

Finally, the DCC principal investigators continue to struggle with issues related to metadata quality and harmonization of different controlled vocabularies in use by the majority of collections contributing item-level metadata.[167]

Over the next two years, the principal investigators expect to integrate collection-level and item-level services as well as customize the interface and metadata design for targeted audiences. At present, the DCC offers two distinct services with separate interfaces: the IMLS DCC Collection Registry and the IMLS Digital Content Gateway. As of January 2006, the registry represents 158 IMLS NLG projects as manifest in 108 primary NLG collection records with 40 additional sub-collection records and 29 associated collections (Cole 2006). Collections are classified according to the Gateway to Educational Materials (GEM) subject schema. As evident from Table 22 below, most collections are assigned more than one subject and contain multiple types of objects. At one extreme, Infomine is assigned to all subjects except Educational Technology and Physical Education. It is also the sole resource classified as Philosophy and all of its sub-categories—Aesthetics, Epistemology, Existentialism, Marxism, and Phenomenology as well as all seven sub-categories of Mathematics.

---

[167] The DCC project Web site provides access to the very useful series of background readings that inform this service's development. Available from, Metadata Roundtable: http://dublincore.org/groups/collections/.

Table 22: IMLS Digital Collections Registry Subject Areas and Object Types (April 2006) N=123 NLG collections plus 40 sub-collections

SUBJECTS	N=163	OBJECT TYPES	N=163
Arts	75	Dataset	6
Educational Technology	9	Image	129
Foreign Languages	5	Interactive Resource	17
Health	7	Moving Image	11
Language Arts	14	Physical Object	46
Mathematics	3	Sound	30
Philosophy	1	Text	112
Physical Education	2	Unknown	2
Religion	8		
Science	26		
Social Studies	131		
Vocational Education	9		

The GEM's classification scheme seems out-of-balance with the subject coverage of the collections in the current deployment of the registry. Around 80 per cent of the collections are classified under "Social Studies" and within this, 104 (or 65 per cent of all collections) are "United States History."

In addition to browsing by subject, users can browse by Object, Place, Title, National Leadership Grant project, and Host Institution. The majority of the collections contain multiple types of images (233) and texts (202). Within these categories, it comes as no surprise that photographs, slides and negatives (105 collections) and books and pamphlets (66 collections) dominate. The registry supports basic and advanced searches. In advanced search mode, users can limit their queries to eight different object types (as noted above). Each entry is linked to the collection's home page, an extensive record about the collection, information about related collections and an annotation about the corresponding NLG project.

Users can link to the Collections Gateway via the "Home" button at the bottom of the screen. (There is no straightforward means to toggle back and forth between the registry and the gateway.) The gateway site supports fielded searches—by Title/Subject/Description, Author/Artist/Creator, Type, Date, and Publisher—deploying basic Boolean operators (AND, OR, NOT) and queries can be limited to all or selected collections (available via a drop-down menu). At present there are 32 collections with item-level records including two that have multiple sub-sets—Heritage Colorado (now Heritage West as discussed in this report) with 22 sub-sets and Museums & the Online Archive of California (MOAC) with 31 sub-sets—for a total of 85 collections altogether.

Users may choose to have results returned in order of relevance. Users can also specify their preference to display all records in short form on one page (up to a maximum of 500); otherwise twenty results are displayed per page. Each entry is linked to its host

identifier; users can view the complete metadata record or add results to a Book Bag and save them (in XML) to a disk. In the left-hand frame of the screen, the search results are summarized according to the collection (and sub-set) to which they are attached. Users can modify or review their search history.

Figure 33: Screenshot of Digital Collections Gateway search for "united states" AND history (April 2006)
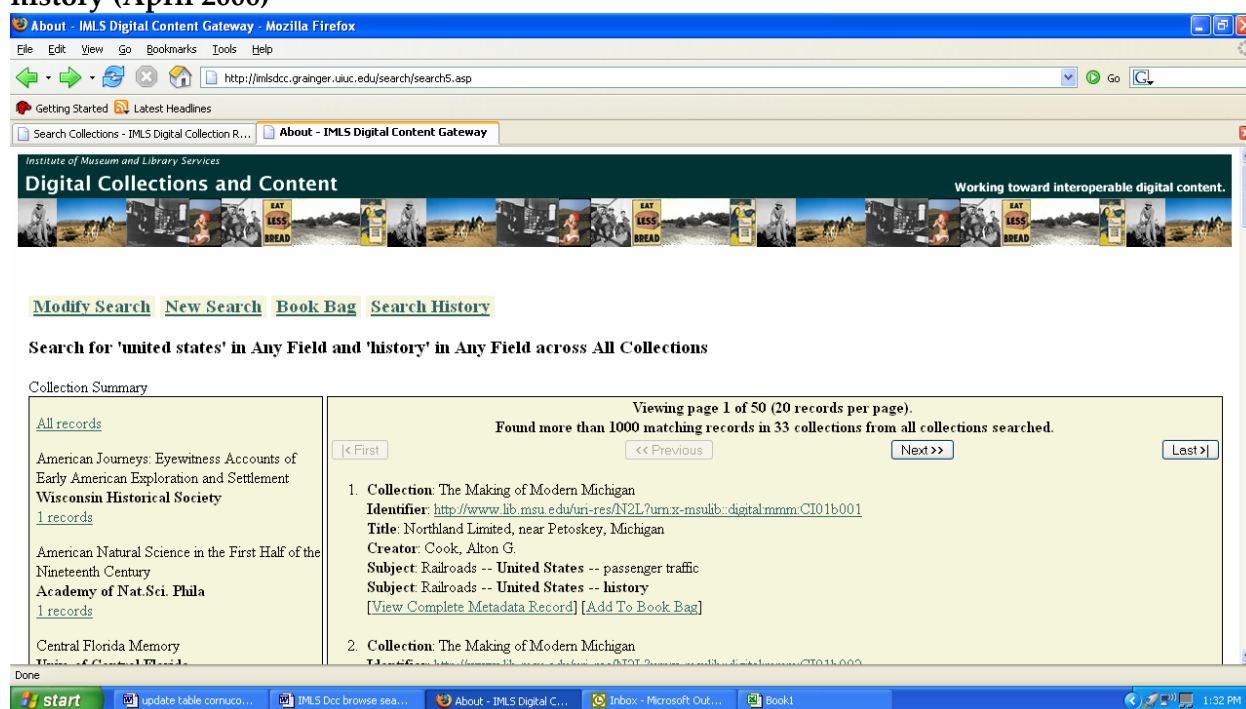
As evident from the table below comparing browsing and searching within and across the registry and gateway, the same query is likely to retrieve different results. In the three examples below, only <human sexuality> retrieves the same collections when using the browse and search features of the registry. However, it returns no hits in the content gateway (HEARTH does not make item-level metadata available). Meanwhile in the case of <dance>, only two collections—the ubiquitous Infomine along with Folkstreams.net—are classified with this GEM subject in the registry, but a search of collections in the registry retrieves a third collection pertaining to the subject, Masterworks Online. Despite its smaller universe of collections, searching the gateway site identifies 85 collections (comprised of 32 collections and 53 sub-sets) with more than 1,000 records relevant to "dance."

Table 23: Comparison of Browse and Search Results in the IMLS DCC Collections Registry and Content Gateway (April 2006)

	DANCE	HUMAN SEXUALITY	UNITED STATES HISTORY
Digital Collections Registry Browse by Subject	• Folkstreams.net • Infomine	• HEARTH (Home Economics Archive: Research, Tradition, and History)	• 105 collections
Digital Collections Registry Search by Subject	Folkstreams.net Infomine Masterworks Online	HEARTH (Home Economics Archive: Research, Tradition, and History)	107 collections
Digital Content Gateway Search by Keyword	44 collections and sub-collections More than 1,000 item-level records None of the 3 collections above included in results	None <sexuality> retrieves 10 records from 3 collections and sub-collections but not HEARTH	33 collections and sub-collections More than 1,000 item-level records

These results illustrate the difficulties that lie ahead as the developers strive to integrate the registry and gateway services into a coherent framework.

4.4.3 DLF Digital Collections Registry

This new registry, maintained by University of Illinois, describes the digital collections hosted or contributed by DLF member institutions and allies that are publicly available and OAI-compliant. As of May 2006, it comprises more than 750 collections from 32 institutions in 19 states plus the British Library (UK). Most of the repository descriptions are based on the collection description schemas that were developed for the IMLS DCC project. In an early stage of development, the site still needs to publicize a collection policy and review its current listings against those criteria. It is accessible from http://gita.grainger.uiuc.edu/dlfcollectionsregistry/browse/.

The DLF Registry has the same user interface as the IMLS DCC Web site, with similar browsing and search options. Browsing collections by time and place reveals that most collections treat late 19th-century and early 20th-century resources about North America. However, the registry embraces collections from ancient to modern times and spans from Africa to South Asia. It is also possible to browse by institution and project. So far only one project is listed—American Culture embracing 46 collections.

Table 24: Number of DLF Digital Collections by Hosting or Contributing Institution (June 2006)

INTERNATIONAL		**MINNESOTA**	
British Library	17	University of Minnesota	5
CALIFORNIA		**NEW HAMPSHIRE**	
California Digital Library	11	Dartmouth College	2
Stanford University	15	**NEW JERSEY**	
University of California-Berkeley	15	Princeton University	12
University of Southern California	26	**NEW YORK**	
CONNECTICUT		Columbia University	18
Yale University	31	Cornell University	65
DISTRICT OF COLUMBIA		New York Public Library	17
Library of Congress	112	New York University	7
NARA	12	**NORTH CAROLINA**	
		North Carolina State University	9
GEORGIA			
Emory University	28	**PENNSYLVANIA**	
Oxford College	1	Carnegie Mellon University	20
ILLINOIS		Pennsylvania State University	15
University of Chicago	27	University of Pennsylvania	23
U of Illinois at Urbana-Champaign	16	**TENNESSEE**	
INDIANA		University of Tennessee	11
Indiana University	17	**TEXAS**	
MARYLAND		University of Texas at Austin	13
Johns Hopkins University	10	**VIRGINIA**	
MASSACHUSETTS		University of Virginia	85
Harvard University	24	**WASHINGTON**	
MIT	30	University of Washington	36
MICHIGAN			
University of Michigan	24	**TOTAL COLLECTIONS**	754

This registry promises to make more visible the digital collections of prominent institutions. Eventually, it should mesh with the DLF Portal (described in section 4.1.8) to offer seamless collection to item discovery and access.

4.4.4 American Memory and Other OAI Digital Collections at the Library of Congress

Update Table 21: American Memory and other OAI Digital Collections at the Library of Congress based on DLF Survey responses, Fall 2005

|	American Memory and Other OAI Digital Collections http://memory.loc.gov/ammem/
ORGANIZATIONAL MODEL	Pilot phase w/ public/private partnership ended; now mainstreamed into LC operations.
SUBJECT	Cultural heritage
FUNCTION	Presents digital content from American Memory, LC Presents: Music, Theater & Dance, Veterans History Project, and Prints & Photographs Online Catalog.
PRIMARY AUDIENCE	Interested public and educators.
STATUS	Established
SIZE	130 collections (30% growth), over 10 million digital items (43% growth), 215,250 OAI-harvestable records (58% growth).
USE	Per day: 200,000 page views and 15,000 searches
ACCOMPLISHMENTS	1. Added 3 million items and about 30 collections. 2. Veterans History Project (VHP) and I Hear America Singing projects demonstrate XML-based approaches. 3. Global Gateway collaborative partnerships with 6 national libraries and other organizations.
CHALLENGES	1. Dealing with multilingual materials in search and display. 2. Creating search across more than one digital conversion project (e.g., American Memory and Global Gateway) 3. Preparation for proposed World Digital Library.
TOOLS OR RESOURCES NEEDED	Tools to support multilingual search.
GOALS OF NEXT GENERATION RESOURCE	Have not yet finalized such goals.

Although the pilot public/private partnership aggregating collections into American Memory has ended, the Library of Congress remains at the forefront in facilitating standards-based digital aggregations and interoperability. In November 2005, Librarian of Congress, James Billington announced LC's campaign to create the "World Digital Library" (WDL) with an initial $3 million contribution from Google (Vise 2005). LC's impressive "Global Gateway" to multilingual resources on world cultures already establishes the precedent of building collaborative digital collections in partnership with other national libraries.

Since the 2003 DLF report appeared, American Memory has grown considerably in size, adding new digitized collections from LC, implementing XML-approaches to audio

projects (e.g., Veterans History Project and the Library of Congress Presents: Music, Theater & Dance), and contributing records to the DLF MODS Portal. Its redesigned front page is a model of clarity and functionality, enabling users to:

- Select a topic to browse collections or link to all browsing options
- Link to the list of all collections
- Look at the day's highlighted collections
- Link for teachers to use American Memory in the classroom (via The Learning Page)
- Submit a reference question to a librarian
- Search all collections
- Link to Help pages
- Read about the history and mission of American Memory
- Contact LC (four different options contingent on the nature of the query)

Figure 34: Screenshot of American Memory's home page
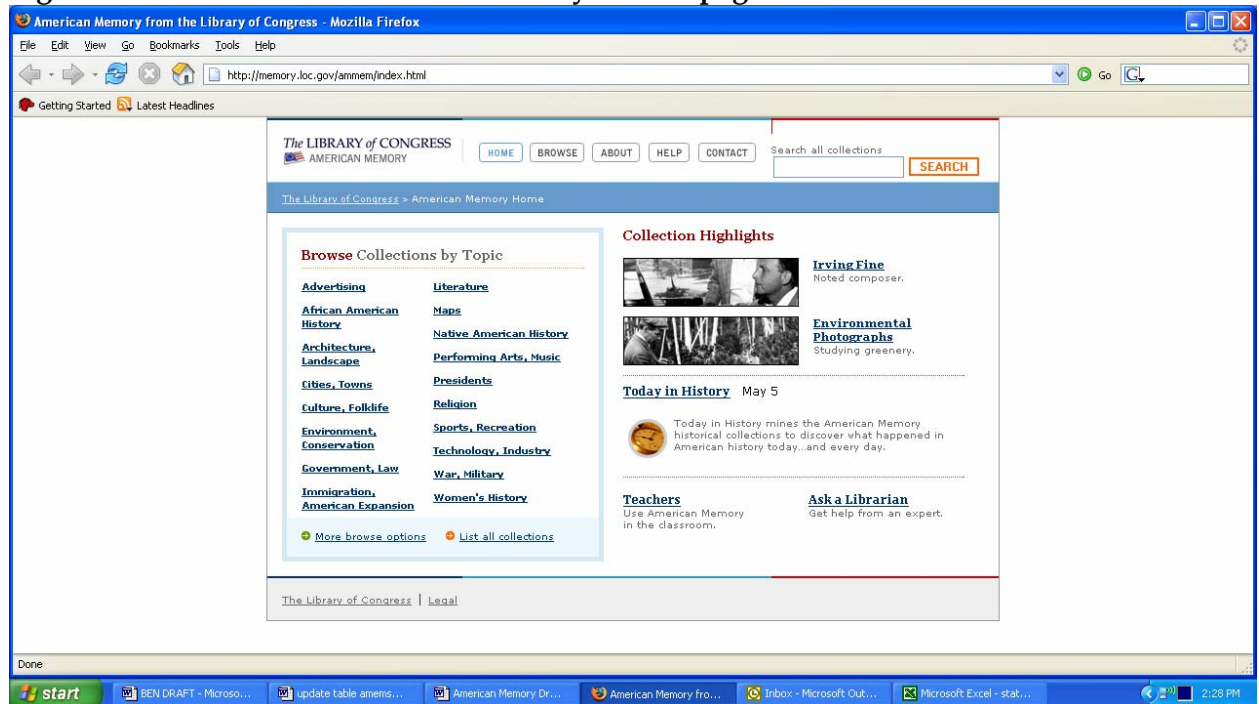
Source: http://memory.loc.gov/ammem/ (May 5, 2006)

In this way, it immediately addresses the varying needs of diverse users ranging from the novice to expert. At the secondary level,

- Users can browse collections by all topics, time period, format, or place.
- The complete list of collections can be sorted by title or subject with the option to view the full collection description.

- The Learning Page aims to serve as the "front door" to American Memory's collections for teachers. In addition to dozens of teacher- and classroom-tested lesson plans, there are featured activities, examples of how to use the collections to develop critical thinking skills, and professional development opportunities, including "self-serve workshops" and tutorials. Teachers can also read "The Source" online newsletter with practical teaching tips, sign up for news alerts about American Memory or participate in monthly live, thematic "chat" sessions (archived transcripts are available as well).

Users can search across all collections or limit their search to specified collections by topic. Results can be displayed in two forms: the default list view (with links to the item and corresponding collection) or gallery view, with clickable thumbnail prints (or when not available, title with link).

An early leader in OAI adoption, LC makes item-level metadata available from American Memory, the Global Gateway and the Prints & Photographs Division's Online Catalog. This includes, for example, all records for LC's moving image materials included in the Moving Image Collections (Johnson 2006). Helpful background documents and guidelines for prospective OAI harvesters are available from the *About* page (see Technical Information).[168] Records are harvestable as sets organized by content type; when more than one set exists, there is the option to harvest individual sets or the combined set. As of May 2006, LC lists the following available records:

- Books (11 individual sets, combined set)
- Ephemera, Pamphlets (1 set)
- Maps, Atlases (1 set)
- Photos (26 individual sets, combined set)
- Posters (2 individual sets, combined set)
- Other Still Visual (4 individual sets, combined set)
- Motion Pictures (1 set)
- Sheet Music (1 set)

A number of aggregation services under review in this report harvest LC records (e.g., OAIster, Perseus, Sheet Music Consortium, American West, MetaScholar, DLF Aquifer, DLF MODS Portal). In addition RLG Cultural Materials (subscription resource) and RLG Trove.net (a free service associated with RLG Cultural Materials) both harvest LC's OAI metadata. LC will utilize the OAI protocol to update a centralized Virtual International Authority File (VIAF) currently under development by OCLC, Die Deutsche Bibliothek, and LC. Intended to serve the international cataloging community,

---

[168] OAI harvesting information is available from
http://memory.loc.gov/ammem/oamh/index.html.

VIAF will include records for Personal Names from selected national libraries (Arms 2003).[169]

4.4.5 Sheet Music Consortium (SMC)

Update Table 22: Sheet Music Consortium based on DLF Survey responses, Fall 2005

	Sheet Music Consortium http://digital.library.ucla.edu/sheetmusic/
ORGANIZATIONAL MODEL	4 Partners: UCLA, Indiana, Johns Hopkins, Duke Also harvest from: LC, Nat'l Library of Australia, Maine Music Box
SUBJECT	Humanities: Music
FUNCTION	OAI aggregator of sheet music.
PRIMARY AUDIENCE	General collection aimed at both general public and academic community.
STATUS	Established
SIZE	110,000 (10% increase)
USE	Per month: 3,909 visits (average)
ACCOMPLISHMENTS	1. Functional development complete & site officially published. 2. Addition of new harvested collections: NLA and Maine Music Box
CHALLENGES	1. Digital collections that are potential targets are not OAI compliant. 2. Incompatible metadata standards
TOOLS OR RESOURCES NEEDED	1. Easy to use software tools that would allow collections to become OAI compliant.
GOALS OF NEXT GENERATION RESOURCE	1. Harvest additional collections. 2. Possible addition of sound recordings 3. Enriched metadata in order to provide better retrieval service.

Intended to leverage the research potential of digital sheet music collections, the Sheet Music Consortium has added two collections—National Library of Australia and the Maine Music Box—to its aggregation since 2003 and is currently adding two more collections from the University of Colorado, Boulder and from the University of Missouri, Kansas City. The Library of Congress and the National Library of Australia have full digital images associated with the metadata records, whereas Indiana University and Duke have a mix of bibliographic metadata and digitized images. For sheet music published after 1922 (and therefore likely under copyright protection), UCLA provides access to the sheet music cover but not the sheet music itself. The Maine

---

[169] Announced by Arms in 2003, an update is anticipated in summer 2006. For background refer to: http://memory.loc.gov/ammem/techdocs/libht2003.html#exploit

Music Box estimates that 62 percent of its collection is in the public domain. For items still under copyright (from 1931 forward), the Maine Music Box does not display images of the score or sound files.

Current data providers are listed below along with the number of metadata records[170]:

- Historic American Sheet Music: Rare Book, Manuscript, and Special Collections Library, Duke University: 20,157
  http://scriptorium.lib.duke.edu/sheetmusic/
- Indiana University Sheet Music: 17,937
  http://www.letrs.indiana.edu/s/sheetmusic/
- The Lester S. Levy Collection of Sheet Music, Johns Hopkins University: 11,590
  http://levysheetmusic.mse.jhu.edu/
- Music for the Nation: American Sheet Music, 1870-1885, Library of Congress: 47,528
  http://memory.loc.gov/ammem/oamh/sheet_music.html
- The Maine Music Box: 11,779
  http://mainemusicbox.library.umaine.edu/musicbox/index.asp
- National Library of Australia Digital Collections/Music: 6,731
  http://www.nla.gov.au/digicoll/
- Digital Archive of Popular American Music (APAM), University of California Los Angeles (UCLA) Music Library: 4,593
  http://digital.library.ucla.edu/apam/

The SMC Web site lists more than 60 institutions that provide some type of public access to digital sheet music collections. Nevertheless, the SMC's aggregation from a mere seven collections contains many more examples of sheet music than other search engines or union catalogs are able to retrieve. SMC could fill a void if it succeeds in attracting more members into the consortium and developing into a full-scale, sophisticated community of practice.

---

[170] Information provided in email correspondence with Curtis Fornadley, Digital Library Architect, SMC (UCLA) on April 14, 2006. The figures at the Web site are out-of-date and incomplete. Some of the April stats are lower than those provided at the native sites. For example, the Maine Music Box reports more than 18,000 OAI-compliant records and the National Library of Australia, more than 9,000.

Table 25: Results for search for "sheet music"

OAIster
>25,000 items from 45 collections
IMLS Content & Collections Registry
>1,000 items from 10 collections
INFOMINE
40 expert-selected resources (does not cite the SMC)
66 robot-selected resources (includes the SMC)
WorldCat
• >20,000 bibliographic records
• >16,000 sheet music with document type: scores
• 40 sheet music with document type: computer files
• 178 sheet music with document type: sound recordings
• 291 sheet music with document type: Internet resources
Sheet Music Consortium
• >110,000 records from 7 collections

Regrettably, SMC does not provide collection descriptions, current harvesting statistics, or details about the number of records, such as those with bibliographic metadata that also have associated digitized images. It is possible, however, to limit searches to digitized sheet music only. The absence of collection-level descriptions is unfortunate since several contributing entities represent multiple special collections from different libraries. For example, the Maine Music Box is an aggregation of five collections drawn the Bagaduce Music Lending Library and the Bangor Public Library.

The Sheet Music Consortium's user interface has not changed since 2003. It supports both basic and advanced searches, including limiting queries to digitized sheet music only. The primary advantage of using the SMC is the ability to search across multiple collections, coupled with the functionally that permits users to select records, add annotations and save (or email) items to a virtual collection that can be shared with others or reserved for personal use.

Levels of Access and Protection in Virtual Collections

The following table shows options for creators of Virtual Collections. Only owners of collections can delete them. Collections without owners will be deleted annually.

Collection State	View	Edit
Private	NO	NO
Public	YES	YES
View / Edit < Password 1 >	Requires Password 1	Requires Password 1
View < Password 1 > or Edit < Password 2 >	Requires Password 1	Requires Password 2

Source: http://digital.library.ucla.edu/sheetmusic/oaihelp.html

Future service enhancements—for example, distinguishing between composers and lyricists, providing access to descriptive elements like plate and publisher numbers, or specifying different types of dates—are hampered by limitations of the available metadata (Davison et al. 2003). As a result, the SMC offers sparse services when compared to the native environments of the constituent collections. Although SMC principals speculate about expanding to include other musical formats, they foresee "a danger in generalizing the service into to [sic] areas that may be better served by other means of discovery" (Ibid). Without any plans to enrich the legacy metadata or integrate SMC more fully into e-learning or e-research environments, SMC seems destined to remain an online union catalog of digitized sheet music with the potential of creating personal or shared virtual collections. While this does fill a need as discussed above, SMC might take a lesson from its partners and review features that they have implemented to develop a more ambitious vision of its future. PictureAustralia, for example, an aggregation that includes the NLA digital music collections, does incorporate different media and also permits discovery by theme.

Figure 35: Screenshot of PictureAustralia (April 9, 2006)

Source: http://www.musicaustralia.org/
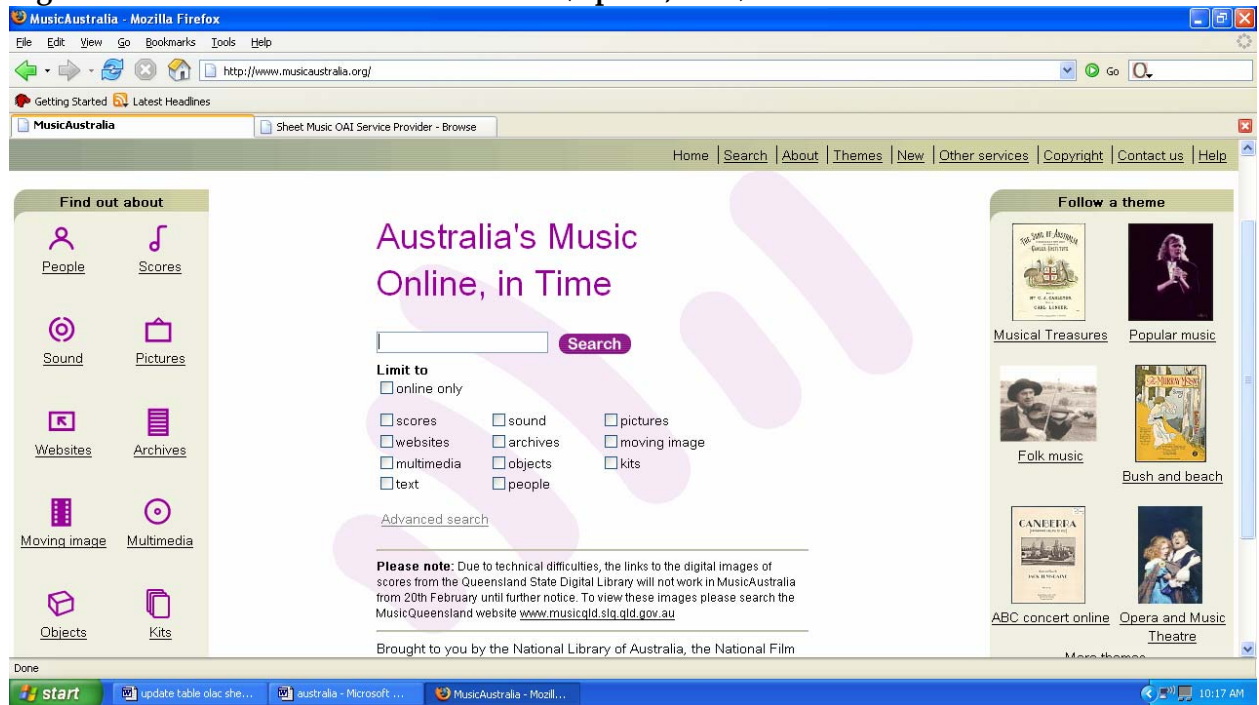
Duke's collection can be browsed by subject content type, illustration type, advertising, and decade (with topical categories). The Maine Music Box offers browsing by subject and sheet music cover art. Moreover, it offers the ability to listen to sound files and has created an instructional module with customized services. Still in its early stage of deployment, its developers believe that "it will take a new generation of music educators to use digital collections as instructional tools." Overall, they "would encourage a vision that provides tools for integrating sheet music collections with other digital libraries," especially promoting their relevance to social and cultural history.[171]

---

[171] Email correspondence with Marilyn Lutz, Maine Music Box, April 19, 2006.

4.4.6 Heritage West (formerly Heritage Colorado)

Update Table 23: Heritage West based on DLF Survey responses, Fall 2005

	HERITAGE WEST (formerly Heritage Colorado) http://www.cdpheritage.org/collection/heritageWest.cfm
ORGANIZATIONAL MODEL	Funded by Colorado Dept of Education, IMLS & NEH.
SUBJECT	Cultural heritage of the western U.S.
FUNCTION	Collaborative efforts of archives, historical societies, libraries & museums located in the western U.S. to make digital collections available to all online audiences.
PRIMARY AUDIENCE	Interested public, educators, researchers, life-long learners.
STATUS	Established
SIZE	77 participating institutions (51% growth); 18 institutional members
USE	[Forthcoming]
ACCOMPLISHMENTS	1. Became a regional collaborative organization. 2. Complete redesign of the CDP Web site in Nov. 2005 w/ user-testing in 2006. 3. Revision/update of CDP Dublin Core Metadata Best Practices document.
CHALLENGES	Sustained funding.
TOOLS OR RESOURCES NEEDED	No response
GOALS OF NEXT GENERATION RESOURCE	Hopes to launch a new interface for delivery of digital content, enabling side-by-side comparison of digital objects, enable the use of METS records, and provide more interactivity for users.

Operating as a not-for-profit with 501c3 status since 2002, the Colorado-based Collaborative Digitization Program (CDP) has expanded its core goals—(1) to achieve high quality digital access to cultural heritage collections and (2) to provide resources and training to create digital surrogates of primary source collections—beyond the borders of Colorado to work with partners across the western United States, including Arizona, Colorado, Kansas, Montana, Nebraska, Nevada, New Mexico, Utah, and Wyoming. CDP members (21 as of April 2006) pay an annual fee ranging from $100 to $2,500 based on their institution's operating and collection budgets. In 2005-06, CDP began to award member institutions with vouchers for free participation in CDP-sponsored workshops or on-site training by CDP staff. CDP carries out its work under the aegis of a Board of Directors, four staff members, and six working groups (Digital Collections, Digital Audio, Digital Imaging, Digital Preservation, Technology, and Metadata).

Figure 36: Screenshot of CDP's Collections Index page

Source: http://www.cdpheritage.org/collection/
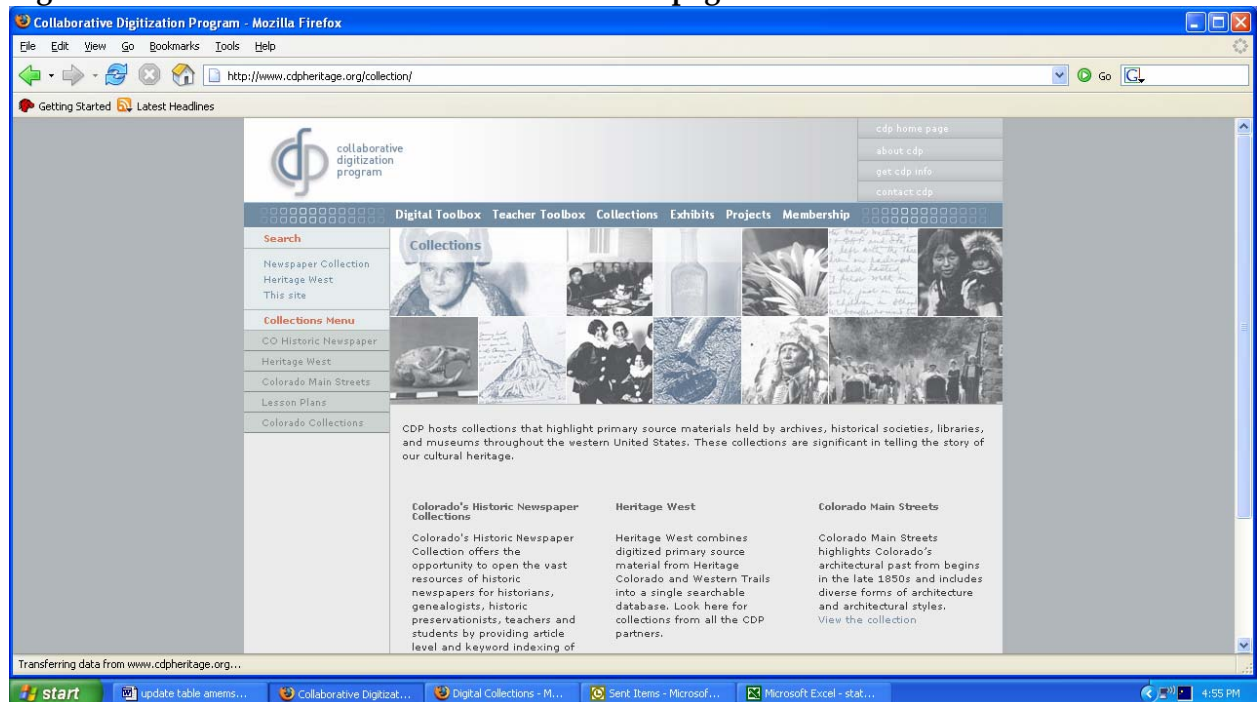
The re-designed Web site offers a multitude of options to meet the needs of varied users from searching CDP's two major collections (Heritage West and Colorado Historic Newspapers) to reading about upcoming CDP workshops, reviewing "Best Practices," linking to "Lesson Plans," or viewing a "Member Spotlight." The "Digital Toolbox" incorporates best practices in digital imaging, Dublin Core metadata, and digital audio; offers information about workshops; and connects to project management guides. "The Teacher Toolbox" is organized into three areas: Why Primary Sources? (links to other primary source sites geared to teachers such as American Memory's Learning Page), Lesson plans, and Professional development.

Heritage West (formerly Heritage Colorado) offers users the ability to conduct unified searches across the digital collections of 77 participating libraries, museums, archives, and historical societies. The new user interface supports basic and advanced searches, as well as searches by topical category. For example, in advanced search mode, users can limit their query to seven collections (comprising the original Heritage Colorado collection, the Denver Public Library and Colorado Historical Society's photographs and images collection, and the five components of the "Western Trails" collection—from Colorado, Kansas, Nebraska, Utah and Wyoming). Search results can be sorted by author, title or date, and saved for emailing. The results are also summarized according to the collection from which they are derived, offering an alternative means of accessing the items.

The Colorado Historic Newspaper Collection (CHNC), CDP's other major database, currently covers 86 newspapers (291,000 digitized pages) published in English, German, Spanish or Swedish in 46 cities and 34 counties throughout the state of Colorado from 1859 to 1928. New material is added on a monthly basis. After extensive user testing, CHNC launched a new search interface in November 2005. It enables users to search newspapers by region within the state and allows them to create a customized group of newspapers for searching. In December 2005, CHNC received a Library Services and Technology Act (LSTA) Continuation Grant from the Colorado State Library that will allow them to partner with the Denver News Agency to run six workshops for educators about the use of historic and current newspaper content in teaching.

As part of the IMLS-funded IMLS Digital Collections and Content gateway, the University of Illinois helped CDP to become OAI-compliant in 2003. OAIster now harvests more than 32,000 items from CDP. While CDP is a collaboration success story, it faces tough decisions about how best to federate searching across multiple databases and whether or not to maintain its own customized software system (DC Builder) or migrate to a commercial solution (Bailey-Hainer and Urban 2004). Reports about CDP are available at its Web site, including a recent presentation by Koelling and Shelstad (2006) summarizing CDP's experience with "Collaborative Digitization Programs."

4.4.7 The American West

Update Table 24: American West based on DLF Survey responses, Fall 2005

	The American West http://www.cdlib.org/inside/projects/amwest/
ORGANIZATIONAL MODEL	Sponsor: William & Flora Hewlett Foundation. Lead institution: California Digital Library. Project partners: CDP (e.g., Heritage West), Harvard, Indiana, LC, Michigan, Virginia, U of Washington.
SUBJECT	Cultural Heritage
FUNCTION	Build a virtual collection on the American West through metadata harvesting and investigate its viability as a tool to assist information resource providers like librarians to better leverage digital content for their specific audiences.
PRIMARY AUDIENCE	Educators
STATUS	Experimental
SIZE	Approximately 250,000 digital objects.
USE	Site not yet released.
ACCOMPLISHMENTS	1. Developing a prototype harvest infrastructure. 2. Ability to ingest metadata-only records into a repository. 3. Concrete steps in developing metadata normalization/enrichment tools.
CHALLENGES	1. Need a better understanding of the needs of audience(s) for OAI-harvested metadata aggregations. 2. Need easier-to-use tools for re-mediating and enhancing harvested metadata. 3. Need clearer use scenarios to drive continued development of OAI aggregation services.
TOOLS OR RESOURCES NEEDED	Widely available metadata normalization tools and tools supporting surfacing topical cohesiveness across highly heterogeneous aggregated collections (could be repository-defined or individual user defined collections).
GOALS OF NEXT GENERATION RESOURCE	This is an R&D project so "next generation" goals are uncertain.

The American West (AmWest) is an experimental project to build a regionally and thematically-focused test bed of OAI-harvested metadata contributed by multiple institutions. Led by the CDL, the AmWest collection has an estimated 250,000 objects contributed by eight partners including the California Digital Library (CDL), the Collaborative Digitization Project, the Library of Congress, Harvard University, and four other university libraries (Indiana, Michigan, Virginia and Washington). Built on the basis of user needs articulated in a series of assessment workshops, AmWest intends to serve a diverse audience ranging from University of California and community

college faculty to academic librarians, K-12 teachers, and public librarians.[172] In particular, it aims to develop tools to configure and integrate virtual collections with local personalized content as well as develop the capacity to deliver learning objects via various platforms such as WebCT.[173]

The project's user assessment reports offer a wealth of insights into the behaviors, needs and expectations of different user groups, while also identifying common ground. Key findings from user interviews resulted in the following recommendations:

- Create **separate gateways** for classroom teaching vs. scholarly research
- Develop **interactive features** to encourage learning and exploration
- Support **advanced search** and filtering
- Allow users to create and publish **personal views** of the collection
- Longer term, encourage users and institutions to **contribute local collections** (CDL 2004)

User input helped to refine the broad topical categories that will form the basis of the site's hierarchical faceted browsing schema. As evident from the Table below, this resulted in numerous modifications to the proposed schema. For example, three categories were added to avoid over-reliance on "Society & Culture" as a "catch-all" category: Family & Community, Leisure & Travel, and Work & Labor. In other instances, categories were revised for precision: Arts became Arts & Architecture, Environment became Land & Resources, and Exploration & Migration became Westward Movement.

Table 26: American West's Broad Topic Categories

Proposed Broad Topic Categories	Revised Broad Topic Categories
1. Agriculture	1. Agriculture
2. Arts	2. **Arts & Architecture (revised)**
3. Business & Industry	3. Business & Industry
4. Education	4. Education
5. Exploration & Migration	5. **Family & Community (added)**
6. Government & Politics	6. Government & Politics
7. Military & War	7. **Land & Resources (revised)**
8. Native Americans	8. **Leisure & Travel (added)**
9. Environment	9. Military & War
10. Race & Ethnicity	10. Native Americans
11. Religion	11. Race & Ethnicity

---

[172] A series of valuable user assessment reports is available from the project Web site http://www.cdlib.org/inside/projects/amwest/.

[173] A list of deliverables is also available at the URL above.

| 12. Science & Technology
13. Society & Culture | 12. Religion
13. Science & Technology
14. Society & Culture
15. **Westward Movement (revised)**
16. **Work & Labor (added)** |

Source: Adapted from Appendix I, Poe 2005, 11
http://www.cdlib.org/inside/assess/evaluation_activities/docs/2005/survey_May2005_report.pdf

The principal investigators have also carried out preliminary work on metadata enhancement to support topical clustering and faceted browsing. Given the extensive amount of pre-processing and human intervention involved in enriching the metadata, they propose that further experimentation—perhaps by the DLF Aquifer Project—is required to determine the optimal balance between collaborative and local responsibilities to facilitate automated classification upon ingest of harvested records and reduce the labor-intensive process of clustering to arrive at targeted topical terms (Landis 2006).

4.4.8 DLF Aquifer

Update Table 25: DLF Aquifer based on DLF Survey responses, Fall 2005

	DLF Aquifer http://www.diglib.org/aquifer/
ORGANIZATIONAL MODEL	Collaboration among subset of DLF membership.
SUBJECT	American culture and life
FUNCTION	To build and test library services that can be integrated into a variety of local environments
PRIMARY AUDIENCE	Academic Community
STATUS	Under development
SIZE	Under development
USE	Site not yet released.
ACCOMPLISHMENTS	1. Receiving strong support from the DLF Board in the form of a dedicated staff member. 2. Developing processes for distributed, collaborative work. 3. Making an implementation plan and beginning to execute it.
CHALLENGES	1. Resources: Difficult for participants to carve out time for this collaborative effort. 2. Diversity of expectations: Participant libraries are interested in emphasizing different facets of the project. 3. Flat organizational structure: DLF is a lean organization, which is both an advantage, allowing the initiative to test the limits of the network and a possible limit.
TOOLS OR RESOURCES NEEDED	Outside funding that would allow dedicated project staff. Support for service model development to evaluate organizational effectiveness and

	to plan for sustainability.
GOALS OF NEXT GENERATION RESOURCE	Experiment with methods of aggregation other than metadata harvesting. Enable "deep sharing," the ability to move digital objects from domain to domain, (e.g., modifying and re-depositing them in a different location in the process.)

Leveraging the quality digital content developed by the Digital Library Federation (DLF) Libraries in American culture and life, the DLF Aquifer is a collaborative project, open to all DLF members, with fourteen current participating institutions, to build an open distributed library. DLF Aquifer will create a test bed of middleware tools and services to support the needs of digital library developers and scholarly end-users alike. To this end, Aquifer has four working groups (Collections, Metadata, Technical Architecture, and Services) along with a coordinating implementation group that sets policy. To date, Aquifer has completed a:

- Business plan:
  http://www.diglib.org/aquifer/AquiferBusinessPlan.pdf;
- Collection development policy:
  http://www.diglib.org/aquifer/Aquifer_CollDevPol_03rev.pdf;
- Descriptive metadata profile
  http://www.diglib.org/aquifer/DLF_MODS_ImpGuidelines_ver4.pdf;
- Functional requirements for metadata harvesting;
- Draft architectural principles;
- Use cases and target audience definitions for services;
- DLF Aquifer Services Institutional Survey
  http://www.diglib.org/aquifer/SWGisrfinal.pdf;
- Prototype metadata aggregation
  http://www.hti.umich.edu/a/aquifer/; and
- Asset action package experiment
  http://rama.grainger.uiuc.edu/assetactions/
  N.B. Link may be unstable due to experimental nature of this site.
  (Adapted from Kott et al. 2005)

The key findings of the institutional survey along with corresponding Aquifer service responses are outlined below:

Table 27: Aquifer Institutional Survey Findings and Service Responses

Key Findings	Aquifer Service Response
Use of digital collections and services is often assessed at point of introduction or update, rather than systematically over time.	Developing an assessment model that can capture the nature of scholarly practice and the long-term integration and use of digital services and resources.
Searching is the most common way that digital collections are used.	Developing tools and services that support meta-searching.
Metadata standardization is the most commonly reported strategy for supporting digital collections	Developing middleware tools that support metadata management activities such as migration, taxonomy assignment, and metadata enrichment.
Institutions and users desire cross-resource discovery tools and greater ability to personalize service options.	Developing tools and services that enable and enhance the integration of digital content into course management systems.
Budgetary, time and personnel constraints challenge the ability of institutions to develop needed services.	Pursuing collaborative collection development to leverage scarce resources. In addition, the other responses above will help to alleviate some of these constraints by supplying models and tools for needed services.

Source: Adapted from DLF-Aquifer Services Institutional Survey Report 2006, Executive Summary: 3-4.

Integral to this report is an annotated list of other user assessment instruments developed by DLF institutions, such as the American West surveys discussed above. These assessment activities are grouped into the following broad categories: Metadata Harvesting and Searching Portals, Collection Aggregation and Display, Navigating and Using Digital Object Collections, and Collecting and Analyzing Usage Data. Together they provide a strong foundation to inform future research about user services in the context of collaborative digital library development.

Three phases are envisioned to roll-out Aquifer service development priorities:

Phase 1: Leveraging institutional infrastructure
- Metadata harvesting (via OAI-PMH)

Phase 2: Enhancing
- Finding (known item/faceted searching via SRU/W)
- Metadata remediation
- Metadata enhancement
- Taxonomy assignment
- Browsing
- Collecting

Phase 3: Deep sharing
- Exporting
- Searching full text
- Integration with course management systems
- Annotation
- Focused crawling

The University of Michigan is hosting the DLF Aquifer portal; it tests out the MODS harvesting for DLF Aquifer collections. As of this writing, the Aquifer prototype Web site contains some 24,000 MODS metadata records contributed by the Library of Congress and Indiana University's Digital Library Program. Eventually, the collection will consist of 250,000 items representing a wide spectrum of media ranging from datasets and images to manuscripts and sheet music. The DLF Aquifer portal is intended to serve as an "administrative" portal, designed as a place for digital library developers to learn more about the DLF Aquifer collections and the richer metadata MODS harvesting provides.

Figure 37: Screenshot of Dublin Core Record from OAIster (General Joshua L. Chamberlain)
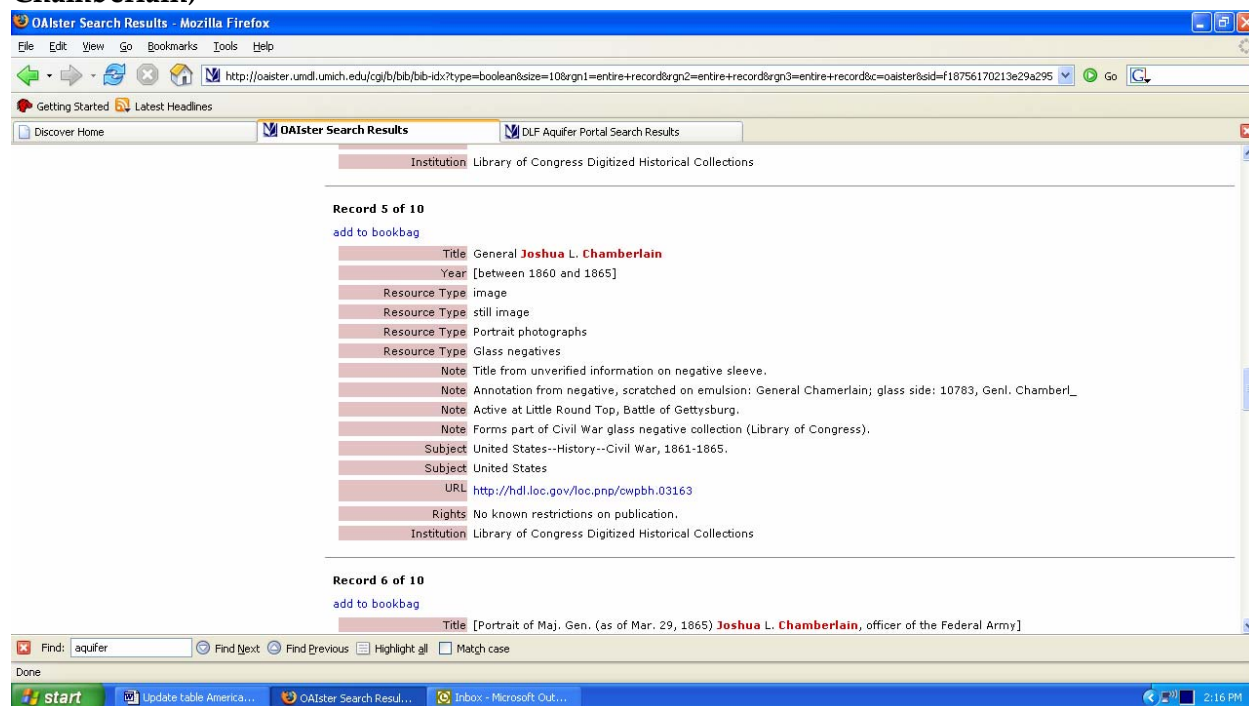
Source: OAIster http://oaister.umdl.umich.edu/o/oaister/ (April 27, 2006)

Figure 38: Screenshot of MODS record from DLF Aquifer for same item
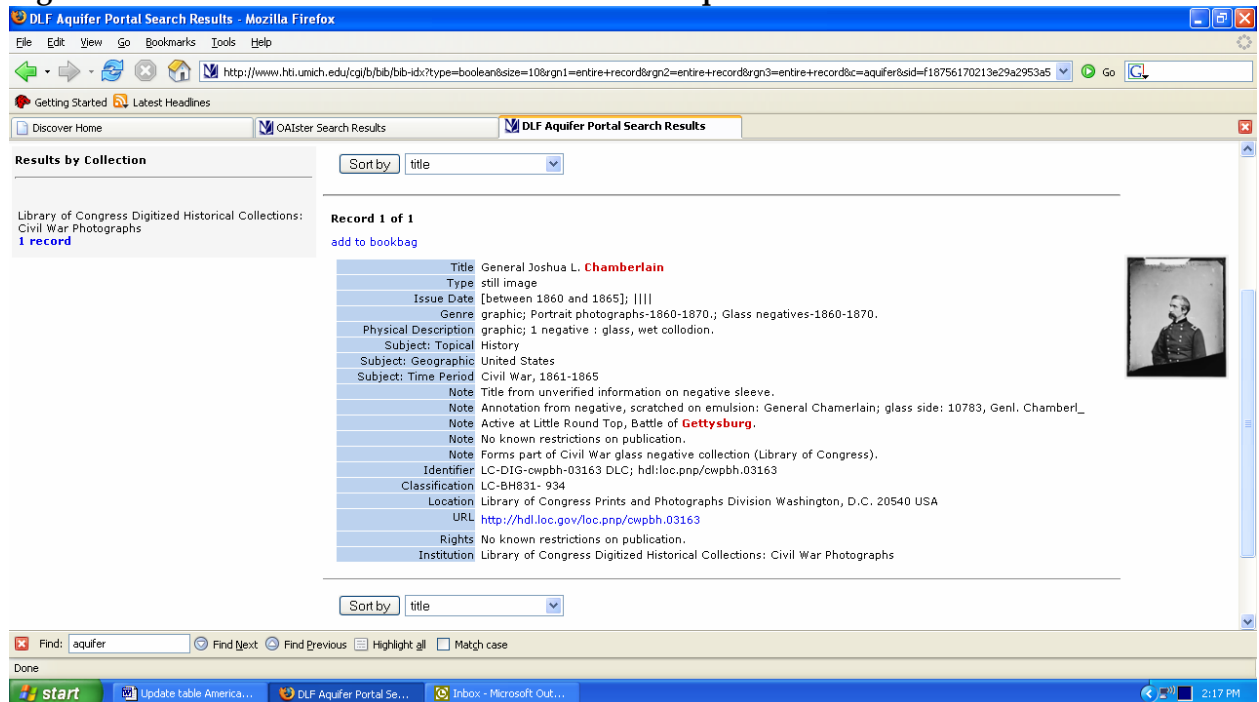
Source: DLF Aquifer http://www.hti.umich.edu/a/aquifer/ (April 27, 2006)

Developed from OAIster, the Aquifer portal features user interface improvements, including thumbnails; additional fields to search (e.g., language and institution); an additional resource type (e.g., dataset) and SRU functionality. Next steps include date normalization and subject clustering. Aquifer is also experimenting with another innovation—"asset action package"—designed "to support a consistent user experience and deeper level of interoperability across collections and repositories" (Kott et al. 2006). This allows multiple views of resources in an OAI context. In practice it enables users to deploy locally-available tools (e.g., for image manipulation, annotation, and saving) with disparately-held content from other repositories that use "asset actions."[174]

---

[174] The experimental deployment of asset action packages at UIUC is available from http://rama.grainger.uiuc.edu/assetactions/index.asp

4.4.9 SouthComb

Update Table 26: SouthComb based on DLF Survey responses, Spring 2006

	SouthComb URL not yet available See http://www.metascholar.org/ for project forerunners.
ORGANIZATIONAL MODEL	Mixed model, currently comprised of grant funding and support from Emory University. Affiliated with Emory University's Robert W. Woodruff Library and the MetaScholar Initiative.
SUBJECT	Multidisciplinary
FUNCTION	Portal for Southern Studies research providing cross-resource search tools that harvest, automatically classify, and meta-search information combined from multiple resources (Web, OAI, and others).
PRIMARY AUDIENCE	Academic Community
STATUS	Under development as of May 2006.
SIZE	Site not yet released.
USE	Site not yet released.
ACCOMPLISHMENTS	This new service builds on achievements of prior work: 1. Refinement of metasearching, semantic clustering, and metadata assignment techniques (MetaCombine and Quality Metrics projects). 2. Development of a conspectus of Southern Studies digital archives. 3. Creation of *Southern Spaces* peer-reviewed Internet journal and its editorial board.
CHALLENGES	1. Metadata format inconsistencies, particularly in describing the resource. 2. Metadata inadequacies, often leading to over-reliance on keyword searching. 3. Sustainability of service: managing the transition from project to ongoing program.
TOOLS OR RESOURCES NEEDED	Metadata normalization tools (some will be developed or deployed in this project).
GOALS OF NEXT GENERATION RESOURCE	As an expansion and improvement on Emory's previous OAI endeavors, SouthComb itself represents a next-generation metasearch service.

Leveraging an impressive series of digital initiatives to advance scholarly communication carried out under the umbrella of "MetaScholar," Emory University received funding in March 2006 from The Andrew W. Mellon Foundation to develop a "capstone" project that would encompass and build on previous metadata harvesting efforts. Tentatively named SouthComb, the project aims to create a sustainable interdisciplinary search portal targeted to Southern Studies research. SouthComb will implement all of the experimental techniques that Emory has developed for harvesting, automatically classifying, and metasearching information from OAI repositories, Web pages, and other scholarly resources. It will establish an advisory panel of Southern Studies scholars at various universities to review and select resources, thus allowing it to

develop sub-portals tailored to meet the needs of particular academic programs. It also aims to widen the cadre of scholars contributing to *Southern Spaces*—an innovative, peer-reviewed Internet journal and scholarly forum created by Emory University.[175] SouthComb intends to advance regional collaboration by advising partner institutions in the use of OAI-PMH tools, such as the Metadata Migrator developed at Emory[176], and by extending participation in the MetaArchive preservation network.[177] To sustain its efforts over the long-term, SouthComb expects to adopt a hybrid business model, consisting of a freely available basic Web resource and a more sophisticated version with advanced features, available on a cost-recovery basis via institutional subscriptions.

According the principal investigators lessons learned and tools developed during previous MetaScholar Initiative projects have contributed significantly to the ability to construct SouthComb.[178] They note the following building blocks:

- semantic clustering and metadata assignment tools developed as part of the MetaCombine work;
  http://www.metacombine.org/overview.shtml
- the multiple resource searching tool developed through the Quality Metrics project;
  http://www.metascholar.org/quality_metrics/
- the institutional collaborative model developed through work on the IMLS-funded Music of Social Change project; and
  http://www.metascholar.org/MOSC/
- scholarly input into pedagogical and research needs for specialized subject areas, as part of earlier Mellon Foundation-funded projects.[179]

As such, SouthComb will emerge as Emory University's next-generation OAI-based service, one that improves the quality of metadata records and the ease of searching and browsing across heterogeneous resources.

User studies conducted for the AmericanSouth project revealed a common desire to search across multiple resources—a finding echoed in other studies of the demands of interdisciplinary research. A hindrance to providing such a tool, however, was the lack of controlled or consistent subject vocabularies for many of these resources. The MetaCombine project, through its focus on semantic clustering techniques, sought to

---

[175] *Southern Spaces* is available from http://www.southernspaces.org.

[176] Metadata Migrator is available from http://metascholar.org/sw/mm/.

[177] For more discussion of MetaArchive, an NDIIPP project, see the project Web site: http://metaarchive.org/.

[178] The remaining description of SouthComb is a slightly edited version of text provided to the author directly from Elizabeth Milewicz, Emory University, in May 2006.

[179] Copies of the final reports for the first MetaArchive project and the AmericanSouth project are posted on the MetaScholar Initiative Web site: http://www.metascholar.org/documents.html.

remedy the obstacles to subject browsing of such heterogeneous materials. In the process the project discovered that focused crawling of Web sites (selectively crawling only Web sites that are relevant to the subject domain under consideration) greatly improved the quality of the results. The harvested results could then be classified according to semantic similarities and organized into taxonomies for easier browsing. These searching techniques, combined with newly developed systems for assigning metadata and visually displaying conceptual connections among records, form the core of the new SouthComb search system.[180]

Other features of the SouthComb search system, developed during the MetaScholar Initiative's Study of User Quality Metrics project, will allow researchers greater precision in identifying resources that are most useful to their work. Using focus-group data on how scholars in the sciences, social sciences, and humanities actually search for and identify quality digital resources for their work, the Quality Metrics project built a new prototype search system that permits scholars to discover resources using both explicit attributes (such as title, author, and other data that currently appear in library records) and implicit attributes (such as citations in journals, usage information from logs, and number of times included in electronic reserves—latent indicators of the scholarly value of a resource). Which resource attributes are highlighted for Southern Studies researchers depends considerably on communications among scholars and the librarians and archivists who provide access to those resources and on focused conversations with scholars about ways they use resources.

The synergistic opportunities continue through the SouthComb portal itself, particularly in its connection with *Southern Spaces*. As a foray into peer-reviewed digital scholarship, the Internet-only journal *Southern Spaces* has re-imagined the possibilities for digital publishing. Through gateways, events and conferences, interviews and performances, and essays that capture the Internet's multimedia potential, the journal's content models the types of scholarly products possible through digital collections and fuels innovations in digital scholarship.

---

[180] Findings from the MetaCombine project are summarized in reports and articles on the project Web site, http://www.metacombine.org/.

Figure 39: Screenshot of an Essay published in *Southern Spaces* that includes video lecture and documentary film footage
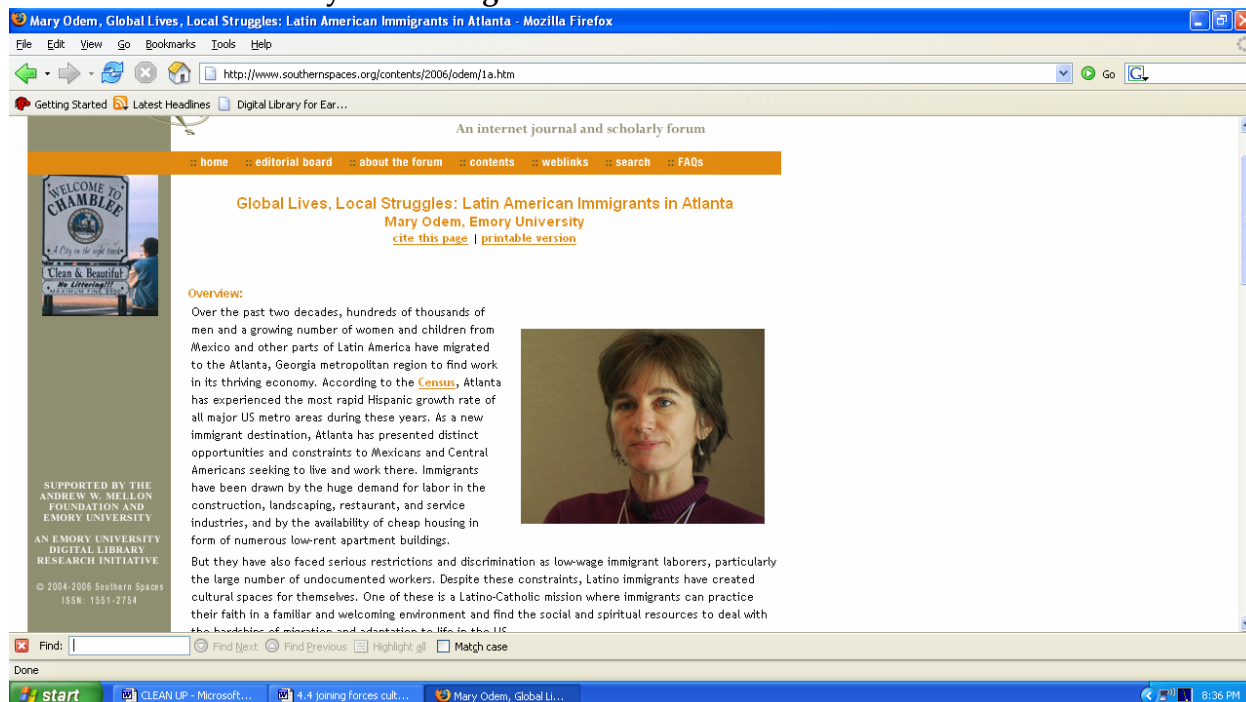
Source: Mary Odem, Global Lives, Local Struggles: Latin American Immigrants in Atlanta. Available from http://www.southernspaces.org/contents/2006/odem/1a.htm

SouthComb hopes to achieve long-term sustainability by providing scholars with the resources they most need and desire. To that end, Emory's digital library has already constructed a Southern Digital Archives Conspectus (SDAC) that describes and provides access to the library- and museum-produced open access digital collections currently available on the topics of history, literature, and culture in the U.S. South from the Colonial Period to the present.[181]

Table 28: Southern Digital Archives Conspectus Classifications (with # of associated collections)

Agriculture and Industry in the American South (37)	Language in the American South (4)
	Law and Politics in the American South (44)
Art and Architecture in the American South (52)	Literature in the American South (29)
	Media in the American South (8)
Education in the American South (61)	Music in the American South (22)
Environment in the American South (32)	Race in the American South (63)
Ethnicity in the American South (32)	Recreation in the American South (20)
Folk Art in the American South (8)	Religion in the American South (22)
Folk life in the American South (17)	Science and Medicine in the American South (11)
Foodways in the American South (4)	
Gender in the American South (29)	Social Class in the American South (20)

---

[181] Available at http://southconspectus.library.emory.edu/.

| Geography in the American South (37) | Urbanization in the American South (26) |
| History, Manners, & Myth in the American South (228) | Violence in the American South (26) |

Source: http://southconspectus.library.emory.edu/SPT--BrowseResources.php (May 2006)

This survey identifies unique collections that would be of great interest to Southern Studies scholars as well as gaps in the digital landscape that could inform future digitization and harvesting efforts. Building SouthComb is conceived as an on-going exercise in community identification and collaboration, leading to greater community investment in digital access to resources, digital scholarship, and digital preservation.

4.4.10 Perseus Digital Library

Update Table 27: Perseus Digital Library based on DLF Survey responses, Fall 2005

	Perseus Digital Library http://www.perseus.tufts.edu/ http://www.perseus.tufts.edu/hopper/
ORGANIZATIONAL MODEL	Tufts University, Classics Depart. w/ NEH, NSF, & other public-private funders
SUBJECT	Humanities
FUNCTION	Evolving digital library of resources for the study of the humanities.
PRIMARY AUDIENCE	Interested public
STATUS	Established
SIZE	1.1 million manually-created and 30 million automatically generated links connect the 100 million words and 75,000 images. 850,000 reference articles provide background on 450,000 people, places, organizations, dictionary definitions, grammatical functions and other topics.[182] N.B.: Corpus comprised of <2,000 texts.
USE	April 2005: served more than 11 million pages to more than 400,000 unique users.[183]
ACCOMPLISHMENTS	1. Rebuilt Perseus text system & released new Web site (Perseus 4.0). 2. Active development of named entity recognition system for historical texts. 3. Improved cataloging of resources including exploration of new standards (MODS, FRBR, etc.).
CHALLENGES	1. Meeting needs of growing audience w/ limited resources, including providing adequate user support. 2. Ability to maintain current services while also implementing research agendas. 3. Implementing a digital preservation strategy.

[182] Information obtained from the "about" page: http://www.perseus.tufts.edu/hopper/about.
[183] Information obtained from the "about" page: http://www.perseus.tufts.edu/hopper/about.

TOOLS OR RESOURCES NEEDED	Exploring various open source tools to support automatic metadata generation, automatic ingestion of digital objects, and improved object relations management systems.
GOALS OF NEXT GENERATION RESOURCE	Eventual release of a named entity browser. Implementation of a distributed editing environment. N.B. Available as of May 2006 at http://www.perseus.tufts.edu/hopper/nebrowser.jsp

In May 2005, the Perseus Digital Library (PDL) released version 4.0 of its system software, facilitating interoperability and closer alignment with Web standards, including support for distributed catalog services based on MODS/MADS/SRU/OAI. The new technology offers capabilities such as:

- Extraction of well-formed XML fragments of primary sources with full TEI-conformant markup permitting developers to create their own front ends;
- Hierarchical (FRBR) catalog (Mimno et al. 2005);
- Discrete XML services for morphological analysis, tables of contents, chunking of larger documents into smaller units, and various categories of searching; and
- Clearer and readily documented API. Tools.[184]

Integral to the new technology plan, PDL ushered in another fundamental change: namely the migration of its core data to the Tufts Institutional Repository—a Fedora object-based architecture better suited to its long-term preservation and access, thereby allowing PDL to concentrate on research and development activities. According to the new arrangement, once PDL research applications have proven their viability, they will move to the institutional repository's production server.

In recent months, PDL has also released significant new content related to 19th-century American documents, taking advantage of new technologies and services (Crane 2006). The user interface permits navigation through texts by "chunking" documents by chapters, parts, pages or tables of contents, and automatically extracts salient places, people and dates for immediate viewing.

At the time of this writing, PDL's content appears to be betwixt and between the original Web site and the new release.[185] It is difficult to correlate the collection overlap (5 at the original site, 4 at the new site) because they are identified by different names.

---

[184] Information available from http://www.perseus.tufts.edu/PR/perseus4.0.ann.full.html.

[185] Original PDL: http://www.perseus.tufts.edu/ and
New PDL: http://www.perseus.tufts.edu/hopper/collections

Table 29: Collections at original and new release of Perseus Digital Library (April 10, 2006)

COLLECTIONS http://www.perseus.tufts.edu	Total Words	Texts	Secondary Sources	Museum Photography	Tools
Classics: Greek, Latin, Archaeology	52,817,833	489	112	166	8
Duke Databank of Documentary Papyri	3,796,476	275			1
English Renaissance: Shakespeare, Marlowe	11,294,934	80	6		
London: Bolles Collection	13,517,917	35			1
American Memory: California	12,799,122	186			
American Memory: Upper Midwest	16,248,751	140			
American Memory: Chesapeake	6,937,628	142			
Tufts History since 1852	771,114	11			
Boyle Papers: History of Science	285,357	47			
	118,469,132	1,405	118	166	10
COLLECTIONS **http://www.perseus.tufts.edu/hopper/collections**					
Classics	46,824,629				
Duke Databank of Documentary Papyri	3,791,687				
Germanic Materials	758,202				
19th-century American	56,140,360				

As evident from the screenshot below, when a particular text is selected, relevant Places, People and Dates are automatically extracted and linked (right-hand frame). The text can also be navigated by chapters and table of contents from the left-hand frame.

Figure 40: Screenshot of *The Writings of John Greenleaf Whittier*, Vol. 1

Source: http://www.perseus.tufts.edu/hopper/ (April 10, 2006)

Developers interested in the evolving technology piloted at PDL should refer to Crane (2006) and related publications at the Web site. In addition to the eventual release of a named entity browser, the principal investigators are researching the implementation of a distributed editing environment whereby users may correct errors, comment on topics, create custom commentaries, user guides, discuss issues with other users, and personalize the Perseus experience.

4.4.11   NINES: Networked Interface for Nineteenth-Century Electronic Scholarship

Update Table 28: NINES based on DLF Survey responses, Fall 2005

	NINES Networked Interface for Nineteenth-Century Electronic Scholarship http://nines.org/
ORGANIZATIONAL MODEL	Sponsored by ALA, ASA, NAVSA, NASSR, SHARP with headquarters at the U of Virginia.
SUBJECT	Humanities
FUNCTION	To provide an online venue for aggregating peer-reviewed scholarly work in British and American literary and cultural studies in the 19th century; to develop a general model for such work; and to facilitate new scholarship using digital tools.
PRIMARY AUDIENCE	Academic community
STATUS	Released to the public in December 2005.

SIZE	The 2005 release aggregates : The Rossetti Archive, The Swinburne Project, Romantic Circles (in part), The Poetess Project, The Walt Whitman Archive, Additional releases are described below.
USE	Not yet available.
ACCOMPLISHMENTS	1. The establishment of the editorial boards and steering committee. 2. The creation of the implementation design for aggregating materials located on distributed institutional servers. 3. The creation of high-level interpretive tools (Collex, Juxta, and IVANHOE) for use within the NINES environment
CHALLENGES	1. Funding to sustain the developing infrastructure. 2. Funding to move paper-based journals that want to become part of the NINES project to online operations. 3. Getting major professional organizations -- in this case, MLA particularly -- to move into active sponsoring mode.
TOOLS OR RESOURCES NEEDED	No response.
GOALS OF NEXT GENERATION RESOURCE	Major goals are to overcome those three major obstacles listed above.

Established in 2004, NINES is a scholarly collective to promulgate peer-reviewed digital scholarship in 19th-century cultural and literary studies, British and American. Headquartered at the University of Virginia under the leadership of Jerome McGann, (John Stewart Bryan University Professor and editor of the acclaimed hypertext project, The Rossetti Archive), NINES is sponsored by five scholarly societies:

- NASSR: North American Society for the Study of Romanticism
- NAVSA: North American Victorian Studies Association
- ASA: the American Studies Association
- ALA: the American Literature Association
- SHARP: the Society for the History of Authorship, Reading & Publishing

More than a dozen other influential humanities computing centers, technical organizations, and digital humanities projects are affiliates.

Guided by a Steering Committee and three domain-specific Editorial Boards, NINES aims to (1) create a shared information management system to coordinate the process of submitting, peer-reviewing and certifying the integration of digital work into NINES and (2) develop a set of customized tools to facilitate knowledge discovery and interpretation (McGann and Nowviskie 2005, 12).

In December 2005, NINES launched a pilot implementation, comprising 24,975 peer-reviewed digital objects, aggregated from six digital projects:

- The Poetess Tradition: http://www.orgs.muohio.edu/womenpoets/poetess/

- The Walt Whitman Archive: http://www.whitmanarchive.org/
- Whitman Bibliography: http://www.uiowa.edu/~wwqr/bibliography.html
- The Swinburne Archiv:
  http://swinburnearchive.indiana.edu/swinburne/www/swinburne/
- The Rossetti Archive: http://www.rossettiarchive.org/
- Romantic Circles Praxis: http://www.rc.umd.edu/praxis/

The June 2006 release adds another 13,000 new objects, incorporating:

- The Charles Chesnutt Digital Archive:
  http://faculty.berea.edu/browners/chesnutt/intro.html
- The British Women Romantic Poets: http://digital.lib.ucdavis.edu/projects/bwrp/
- The Ambrose Bierce Project: http://www.biercephile.com/
- Romanticism on the Net: http://www.ron.umontreal.ca/
- Victorian Studies Bibliography: http://www.letrs.indiana.edu/web/v/victbib/
- The Blake Archive http://www.blakearchive.org/
- Collective Biographies of Women: http://etext.virginia.edu/WomensBios/

A later release, scheduled for fall 2006, will bring in another 30,000 new objects from: The Whistler Correspondence and The Dickinson Electronic Archives. Consultation is underway to integrate other online resources into the aggregation. All contributions are vetted through NINES editorial apparatus prior to their release. The technology described below, enables users to browse and search collections; registrants can collect and annotate selected search results. http://www.emilydickinson.org/ http://www.whistler.arts.gla.ac.uk/correspondence/

The NINES technology plan has evolved from a centralized, hierarchical approach requiring compliance with a monolithic set of governing standards for text markup, metadata, interface, and archiving to a more flexible, collaborative and non-hierarchical design, relying on RDF (Resource Description Framework) syntax to facilitate description and semantic integration of NINES resources. NINES uses a customized open-source indexing system and the Lucene search engine, customized to integrate faceted browsing. COLLEX, a tool developed by NINES, serves as the backbone of the system that "brings this indexing and search design framework into a collaborative research environment" (Ibid, 15). COLLEX "leverages current developments in folksonomy and semantic-web technology to perform data mining operations and enhance knowledge discovery," . . . leading scholars and students "to see connections among digital objects, based on the contexts into which those objects have been placed (implicitly or explicitly) by past scholarly activity in the system" (Ibid, 15-16). Users of COLLEX can:

1. collect, tag, analyze, and annotate trusted objects (digital texts and images vetted for scholarly integrity);
2. reorganize and publish objects in fresh critical perspectives;
3. share these new collections with students and colleagues, in a variety of output formats; and,
4. without any special technical training, produce interlinked online and print exhibits using a set of professional design templates. (Ibid, 16)[186]

Two screenshots from the COLLEX prototype appear below. The first presents the initial view into the NINES browser with COLLEX sidebar incorporated. Users see the featured NINES exhibits at the top of the screen, have access to the most popular folksonomy tags, and have access to faceted browsing and search. The second screenshot is a view of the system after some constraints have been introduced in the faceted browser. More specifically, it depicts browsing a user-created tag ("reflection") in the sidebar.

Figure 41: Screenshot of NINES browser with COLLEX sidebar
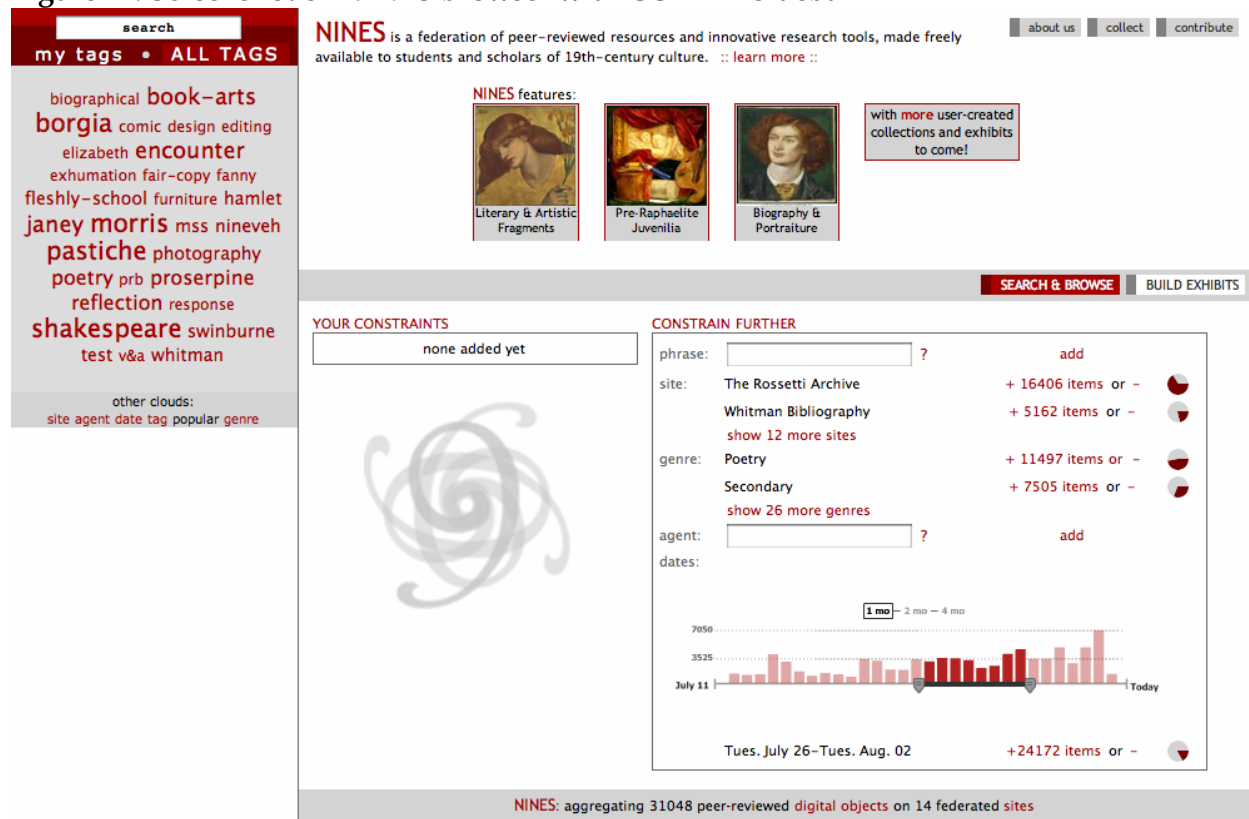
Source: Screenshots provided by NINES' developers (May 8, 2006)

---

[186] For a full description of COLLEX see Nowviskie 2005; available from http://www.nines.org/about/Nowviskie-Collex.pdf

Figure 42: Screenshot of NINES browser with user-tag "reflection" introduced

Source: Screenshot provided by NINES developers (May 8, 2006)

In addition, NINES has developed two other interpretative applications that can be used in tandem with COLLEX or independently. Juxta is a collation and text comparison tool with analytical visualization capability (released for testing in February 2006). IVANHOE is a collaborative interpretative play space, especially designed for pedagogical use[187] (McGann 2005).

NINES aims to increase participation in digital scholarship by awarding competitive fellowships during the summer to train scholars who are developing digital projects and by working with journal editors to facilitate the migration of paper-based journals to digital or hybrid formats. As of this writing, NINES is seeking grants to extend its development for another two years and also hopes to gain endorsement from the Modern Language Association of America.

---

[187] More information about IVANHOE is available from http://www.patacriticism.org/ivanhoe/index.html.

4.4.12 Current Issues and Future Directions

- Synergies among these services are apparent in terms of sharing collections (Heritage West/American West; DLF Aquifer/American West/Southcomb), metadata schema (Cornucopia/IMLS DCC), user interface and search systems (IMLS DCC /DLF Digital Collections Registry), and tools development (American West/DLF Aquifer/SouthComb). In these and other ways, this cohort collaborates and builds on each other's work, directly or indirectly. They represent, however, different models of achieving organizational sustainability.

- The more complex services, such as DLF Aquifer, SouthComb, and NINES must engage at least three different communities of practice: scholarly and disciplinary circles; digital library technical domains; and e-learning and/or e-research service communities. For these aggregation services to flourish over the long-term, they have to be cognizant of the needs and trends across all three sectors. NINES, for example, garnered support from a host of relevant scholarly societies, humanities digital computing centers and other digital libraries in addition to establishing editorial boards charged with peer-review oversight for content. While it is scholar-led, the service itself is embedded in a library setting at the University of Virginia. DLF Aquifer, on the other hand, operates in a decentralized manner where participants agree to terms spelled out in a business plan, collection development policy and other technical specifications. It is driven by the DL community and loosely informed by scholars, with the intention of building prototype tools and services that can be applied at different institutions to meet their particular needs. SouthComb has a multi-faceted organizational structure, with scholars taking the lead for some components, such as the journal *Southern Spaces*, while the library develops the tools and finding system. The library and scholars work in tandem to identify new content and bring it into the aggregation. A challenge facing all three of these collaborations is how to achieve a reputation of sufficient stature that other scholars and libraries are willing to contribute their time (for example, peer review or tool development), content, scholarship, and financial resources outside their local institutional setting. In short, are the benefits of collaboration, fruits of cooperative labor, and reward system adequate to carry the day?

- It is important to acknowledge that virtually all of the services under review in this section play a major role in empowering other data providers to achieve interoperability through OAI implementation and promulgation of best practices. Projects like Cornucopia, IMLS DCC, and Heritage West provide constituent services with the tools they need to maintain control over their own information environments while also fostering their ability to contribute to aggregations. In this way, they have helped to increase the quantity and quality of data providers.

- Representatives from many of these services (e.g., IMLS DCC, American Memory, American West, SouthComb, DLF Aquifer) are directly involved not only in developing the "Best Practices for OAI Data Provider Implementations and Shareable Metadata," (a joint DLF and NSDL initiative), but also in creating the

means to achieve consistent and adequate metadata. Nevertheless, many of them note the unmet challenge of having sufficient automated and semi-automated tools at their disposal to enhance and remediate metadata for scholarly use. A particular challenge and focus of activity among these services is devising methods to achieve subject classifications, thematic groupings or topical clustering across large, heterogeneous collections. In July 2006, the Digital Library Federation and The Andrew W. Mellon Foundation are sponsoring "The Metadata Enhancement and OAI Workshop" at Emory University where DL specialists will examine automated and semi-automated strategies for metadata enhancement and remediation scenarios involving the OAI protocol. Some of the scenarios being considered include normalization of date and format fields and taxonomy generation/assignment. The workshop will result in an agenda for specific experiments to assess various scenarios collaboratively, especially as part of the DLF Aquifer project.

- Other common challenges revolve around standardization of terminology, multilingual metadata and search support, and aligning collections with their associated items in meaningful contexts.
- Future generation plans include new user interfaces that enable side-by-side comparison of documents, more interactive features and interpretative tools, the capacity to move complex digital objects from domain to domain, and the ability to migrate core data to production sites where preservation services are also available.
- Finally, funding strategies (aside from grand and foundation funding) to ensure long-term viability is common concern for these services. Heritage West has a membership fee structure. DLF Aquifer's business plan includes a provision to consider fee-for-service components after its initial development and SouthComb will make some of its services available through institutional subscriptions. Ultimately, the longevity of this cohort rests on how well it meshes with pedagogical practices, e-scholarship, and lifelong learning pursuits.

4.5 User Alchemy: Discover, Deliver, Divine

First Choice for Information—by College Students across all Regions
"Which source/place would be your first choice?

Search engines	72%
Library (physical)	14%
Online library	10%
Bookstore (physical)	2%
Online Bookstore	2%

Source: OCLC 2006, A-20.

The services under review in this section are all attempting to distinguish themselves from generic but hugely popular search engines by customizing their approach to meet the needs of the academic community. From niche search engines to customized and "accessorized" portals, they are components of evolving finding systems that move

beyond discovery to the delivery and re-use of digital content. They represent a progressive spectrum of solutions to integrating search results from federated searching to metasearch systems, differentiated by "just-in-case processing" versus "just-in-time processing" (Sadeh 2006).

The review starts with Scirus, a federated search service which has increased its coverage significantly since 2003, by extending Web crawling and indexing to a much broader array of subjects. It moves on to Infomine, a collaboratively developed index and catalog of expert-selected and robot-retrieved Internet resources. Next, a service from the UK, "Intute," along with the transition to this new name, hopes to become a trusted, first-choice, Internet "mentor" and filter for quality information. More than a search engine, Intute is embedding its resources and services into a variety of teaching and research environments. Finally, the California Digital Library's Metasearch Initiative represents a coordinated and multi-faceted digital infrastructure that integrates all resources—irrespective of origin, host location or protocol—into user-controlled service environments. The later two projects, not yet fully deployed, show the promise of how various standards and best practices come together in service to the academic community.

4.5.1 Scirus

Update Table 29: Scirus based on DLF Survey responses, Fall 2005

	Scirus for Scientific Information Only http://www.scirus.com/srsapp/
ORGANIZATIONAL MODEL	Elsevier
SUBJECT	Sciences [and other scholarly information]
FUNCTION	Multidisciplinary search engine, focusing on science
PRIMARY AUDIENCE	Research Community
STATUS	Established
SIZE	Crawls over 217 million science-related pages, consisting of 179 million Web pages, as well as 38 million records from both proprietary & OAI-compliant sources (including journals, institutional repositories, patents, e-prints from arXiv, technical reports from NASA, etc.)
USE	Per day: > 115,000 searches.
ACCOMPLISHMENTS	1. Significant increase in the size and variety of content types in Scirus's index. 2. Improvements in indexing process and content classification 3. Improvements in user interface
CHALLENGES	No response
TOOLS OR RESOURCES NEEDED	No response
GOALS OF NEXT	No response

GENERATION RESOURCE	

Scirus[188]. Elsevier's award-winning search engine[189], continues to grow in content, types of information, and functionality. "How Scirus Works" (updated in August 2004) describes its process of gathering and classifying data into its index; it also explains search functionality, ranking and search refinement.[190] Scirus uses a combination of focused Web crawling, based on a "seed list" of URLs manually checked for scientific content, and database loads from its partners (e.g., ScienceDirect, MEDLINE, LexisNexis) and OAI harvesting (e.g., from arXiv.org, RePEc, NDLTD). As of early March 2006, it boasted more than 250 million Web pages with the majority derived from educational institutions; the slowest growth in representation is from commercial sites.

Table 30: Scirus Web Page Counts by Domain (March 17, 2006)

	August 2003	March 2006
.edu sites	45 million	83 million
.com sites	18 million	22 million
.org sites	14.8 million	25 million
.ac.uk sites	5.5 million	10 million
.gov sites	4.7 million	6.5 million
Other STM and university sites around the world	Over 40 million	Over 68 million

Source: Scirus "about" page.

Since 2003, Scirus has augmented considerably its journal and "preferred Web sources" content from a combination of subscription-based (e.g., Crystallography Journals Online, Institute of Physics, Scitation) and freely available OAI sources (e.g., PubMed Central, DiVA, MIT OpenCourseWare, NDLTD, RePEc). (It has also dropped several sources including Beilstein Abstracts and its own—Elsevier—Chemistry, Mathematics and Computer Science Preprint Archives. These are available on a subscription basis via Chemweb, Elsevier's Chemistry portal and other Elsevier portals.)

---

[188] This description concentrates on changes to Scirus occurring since the 2003 report. It makes no attempt to cover all of Scirus's features. Readers seeking a more thorough description of features should read the Scirus White Paper (see footnote below) or Gerry McKiernan's E-profile of Scirus (2005).

[189] Scirus received the "Best Directory or Search Engine Website" Web Award from the Web Marketing Association for the second consecutive year in 2005. It was awarded the "Best Specialty Search Engine" by Search Engine Watch in 2001 and 2002, and received an honorable mention in 2005 for this category. See September 2005 press release available from http://www.scirus.com/press/pdf/webaward.pdf .

[190] Scirus White Paper, "How Scirus Works," updated August 2004; available from http://www.scirus.com/press/pdf/WhitePaper_Scirus.pdf.

While its coverage is strongest in the sciences (especially its journal sources), Scirus's subject scope now expands across all disciplines due to the inclusion of multidisciplinary content from ETDs, academic OAI repositories, and broader Web crawls. The category below, "Digital Archive" currently consists of records from Organic eprints and the UMDL. Elsevier expects this category to expand significantly in the course of the year. Scirus now also indexes news sources and offers news results in a dedicated section at the bottom of the results page. The feature includes news items form the last 30 days and ranks results by relevance and date. Up-to-date news from the New Scientist is also available directly from as a link off the home page.

Table 31: Titles and Record Counts of Scirus's Proprietary and OAI Sources (March 17, 2006)

Journal Content with Number of Full-Text Articles (or Citations)	**Preferred Web sources** (e-prints, technical reports, ETDs, patent data, course materials)
BioMed Central: 6,515 Crystallography Journals Online: 56,310 Institute of Physics: [207,000] MEDLINE/PubMed: 15.2 million citations PubMed Central: 285,500 Project Euclid: 28,510 ScienceDirect: 5.6 million Scitation: 318,760 SIAM (Society for Industrial & Applied Mathematics): 7,300	arXiv.org: 311.065 Caltech: 3,058 DiVA: 1,500 Cogprints: 2,175 MIT OpenCourseWare: 33,050 NASA: 12,265 NDLTD: 149,381 Patent Offices data from esp@cenet (European Patent Office) and the US Patent and Trade Office or via LexisNexis platform: 13 million RePEc: 163,800 University of Toronto T-Space: 2,080 Digital Archives[191]: Organic eprints [4,360], UMDL (University of Michigan Digital Library) [198,000]

Source: Sources are from search categories available from the Advanced Search page. Record counts are from the "about" page. Figures in brackets and additional information about patent sources are from email correspondence with Sharon Mombru on March 17, 2006.

Users can perform a search within or across three broad categories: all journal sources, preferred Web sources, or other Web sources. (Scirus indexes Web pages and their relationships, classifying the content by subject and information type through utilization of a collection of dictionaries with more the 1.6 million scientific terms, pattern recognition tools, and linguistic analysis. This enables users to limit searches by eight different information types and twenty subject areas as well as six file format types.

---

[191] The constituent contents of "Digital Archive" do not appear on the Scirus advanced search or "about" page. Information obtained in email communication from Clive Clarke on February 22, 2006. He noted that more sources may be added in the future.

Table 32: Scirus Delimiters: Information Types, File Types, and Subjects (March 17, 2006)

Information Types	Subjects
Abstracts	Agricultural and Biological Sciences
Articles	Astronomy
Books	Chemistry and Chemical Engineering
Company homepages	Computer Science
Conferences	Earth and Planetary Sciences
Patents	Economics, Business and Management
Preprints	Engineering, Energy and Technology
Scientist homepages	Environmental Sciences
	Languages and Linguistics
File Types	Law
PDF	Life Sciences
HTML	Materials Science
Word	Mathematics
Postscript	Medicine
TeX	Neuroscience
PowerPoint	Pharmacology
	Physics
	Psychology
	Social and Behavioral Sciences
	Sociology

Source: Scirus Advanced Search page

Searches can be narrowed to particular authors, journals or titles and restricted to specified date ranges. A sample query for journal articles published in the "Institute of Physics" with the keyword "laemmli," returns results with the search term highlighted (in this case, it appears among the article's references) and clearly indicates the source of the published article.

Figure 43: Sample search for "laemmli" restricted to IoP journal articles

Predicting the function of eukaryotic scaffold/matrix attachment regions via DNA mechanics
Ming Li / Zhong-can Ou-Yang, *Journal of Physics: Condensed Matter*, Aug 2005
...near future to elucidate how universal our hypothesis is. References [1] Freeman M 2000 Nature 408 313 [2] Paulson J R and **Laemmli** U K 1977 Cell 12 817 [3] Phi-Van L and Str¨atling W H 1988 EMBO J. 7 655 [4] Levy-Wilson B and Fortier C 1989 J. Biol. Chem... **Published journal article**

available from IoP
view all 3 results from Institute of Physics Publishing
similar results

Scirus automatically performs "intelligent query rewrites," suggests "did you mean?" queries, and lists alternative keywords to refine or expand searches. Search results are returned according to relevance ranking (determined by an algorithm that takes into account word location and frequency as well as the number of links to a page) or date.

Users can refine, customize or save searches and email of export selected search results to their reference management application.

An Advanced Search for:
        EXACT PHRASE <avian influenza>
        AND
   ALL THE WORDS <pandemic>

Is automatically rewritten as a Basic Search for:
        <"avian influenza" AND (pandemic)>

It retrieves 14,769 total results including 595 journal results, 29 preferred Web results, and 14,145 other Web results. Terms to refine the search are located in the right-hand margin. Several sponsored links follow from commercial suppliers of products for avian flu protection.

Figure 44: Screenshot of Scirus Search Results Page (March 19, 2006)
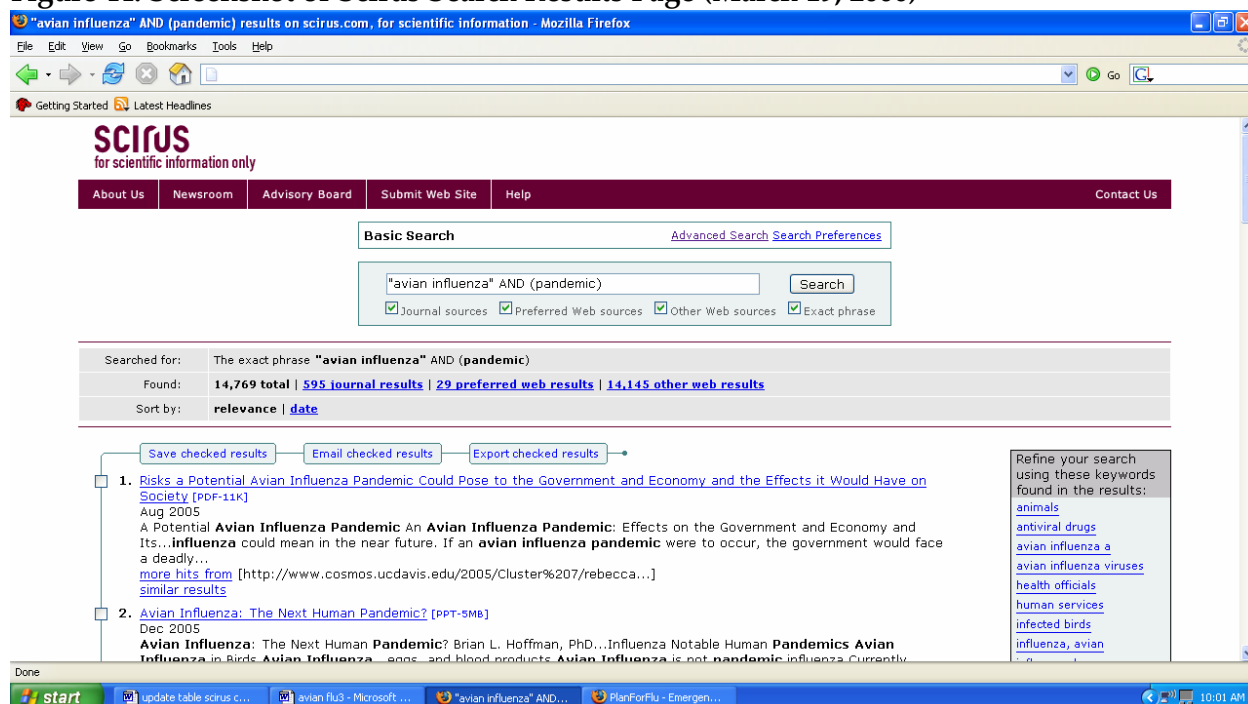
According to Elsevier representatives, "Scirus indexes sources of STM-relevance in the broad sense of the world—scientific, technical, medical, social sciences, etc."[192] With its more expansive subject scope, Scirus may need to revise its qualifier from "for scientific information only" to "for scholarly information only." The Scirus toolbar and customizable search query boxes (for general searches or limited by subject and other

---

[192] Email correspondence with Sharon Mombru on March 17, 2006.

fields) can be added to external Web sites. The Scandinavian aggregator of academic repositories, DiVA, (which is harvested by Scirus) for example, offers users three search options at its Web site, including the ability to use the Scirus search engine (restricting the query to DiVA content, preferred Web sources, or all of the scientific Web).[193]

4.5.2 INFOMINE

Update Table 30: INFOMINE based on DLF Survey responses, Fall 2005

	INFOMINE http://infomine.ucr.edu/
ORGANIZATIONAL MODEL	UC-Riverside and national network of libraries w/ IMLS and NSDL funding
SUBJECT	Multidisciplinary
FUNCTION	Virtual library of expert and machine-gathered scholarly Internet resources
PRIMARY AUDIENCE	Academic community
STATUS	Established
SIZE	210,000 resources (110% increase) of which an estimated 17,000 have associated full-object representation
USE	Average successful requests for pages per day: 2,190. Per month: 66,795. Per year: ~800,000
ACCOMPLISHMENTS	1. Populating the database with robot records, mostly created from the iVia virtual library crawler and machine-generated metadata (using iVia classifiers). 2. Using new versions of iVia open source Software that has increased the accuracy of its classifiers and focused crawlers. 3. Collaborating and sharing metadata with other projects.
CHALLENGES	1. Continued funding of programmers and metadata specialists. 2. Sustaining an active level of participants in the INFOMINE collecting cooperative in various subject areas. 3. Increasing the level of expert & robot records in the collection of INFOMINE Scholarly Internet Resources.
TOOLS OR RESOURCES NEEDED	New versions of improved classifiers and crawlers to help scale with the increase of scholarly Internet resources.
GOALS OF NEXT GENERATION RESOURCE	1. More customizable features for INFOMINE users. 2. Expanding subject areas. 3. Harvesting and sharing metadata with other digital and virtual libraries. 4. Providing more, rich full-text. 5. Continue to improve the iVia open source software.

---

[193] See http://www.diva-portal.org/scirus.xsql?lang=en.

As reported in 2003, INFOMINE is a national collaborative project led by the University of California-Riverside (UCR) to create a virtual library of scholarly Internet resources, utilizing the open source iVia software platform. INFOMINE provides access to more than 200,000 freely available and commercial resources organized into nine subject areas and covering a wide spectrum of media formats. The INFOMINE database represents a hybrid approach to collection building, relying on a combination of contributions from library subject specialists and focused Web crawling. As a result, searches can be limited to "expert-selected" or "robot-selected" items. Expert-selected content constitutes less than 20 percent of the total content.[194]

The flexible, modular system is designed to facilitate cooperative collection-building of the centralized database while also providing institutions with the tools they need to develop customized Internet resource discovery systems with local branding and incorporation of proprietary materials. *MyInfomine* (also known as *MyI*) supports building sub-collections of INFOMINE and enables contributors to create *MyInfomine Categories*, add records to these categories, and perform searches on them.[195] For example, librarians at the University of California-Riverside have created *MyI* categories to create course-specific Internet resource guides as well as to track medical indexes and databases.

iVia (http://ivia.ucr.edu) is an open source system for automatically and semi-automatically building library-related metadata and rich text collections of Internet available resources. Web-based virtual libraries like INFOMINE, subject portals and catalogs benefit. The codebase (250k lines of C++) has been designed by and for librarians and computer scientists at the University of California, Riverside, the NSDL< Library of Congress, Cornell University Library and others. The goal of the software is to amplify expert effort in collection building and foster collaboration.

iVia supports automated metadata generation to assign Library of Congress Subject Headings and LC Classifications to resources (Mitchell et al. 2004, Paynter 2005). Building on iVia, Data Fountains, currently under development, identifies itself as "a national, cooperative information utility for shared Internet resource discovery, metadata application and rich, full-text harvest of value to Internet portals, and library catalogs with portal-like capabilities." It uses expert-guided and focused crawlers supported in semi-automated (requires expert refinement) and fully automated modes (Mitchell 2006). The expert (or manually) guided crawler drills down from a given URL, whereas the focused crawler uses techniques of co-citation and similarity analysis to identify intensely interlinked and high value resources in a subject. The "Nalanda iVia

---

[194] A statistical table of Expert Created Content broken down by subject and type of resources is available from
http://infomine.ucr.edu/about/content.shtml.
[195] As described on INFOMINE's Research & Development pages, available from
http://infomine.ucr.edu/projects/integration/.

Focused Crawler" features an "apprentice learner" program that enables it to follow the most promising links; crawling is also improved by utilizing a combined HITS and PageRank algorithm (Ibid).

In January 2004, iVia received a two-year sub-contract from NSDL to integrate iVia software into NSDL's Core Integration efforts. This has enabled NSDL to harvest item-level metadata from iVia's server for selected NSDL collections that did not include detailed metadata. By October 2004, NSDL reported that they successfully submitted a URL to iVia's Expert Guided Crawl Service and reviewed the results to delete inactive or irrelevant sites, then harvested the metadata using OAI.[196] Phipps, Hillmann and Paynter (2004) discuss NSDL's service interaction with INFOMINE, enabling "loosely-coupled third party services to provide metadata enhancements to a central repository."

INFOMINE's advanced user interface supports searches that can be restricted by fields (author, title, subject, keyword, description, full text, and MyInfomine) or by broad subject area as well as restricted by source (expert-created or expert- and robot-created), access (all, free, or fee-based), and type (e.g., article databases, datasets, patents, preprints and working papers, etc.). Users can select the length, number, and order of the results' display. In addition they can browse within several indexes—including LC subject headings and classifications—by keyword, author, title, or what's new (entries added in the last 20 days). Although users can submit comments about resources, there are no other post-processing functions such as saving, downloading, or emailing search results.

INFOMINE's search tips are exemplary and include a succinct, yet extensive review of how to combine Boolean and proximity operators.

Table 33: Combining Boolean and proximity operators in INFOMINE

Search Statement	Executed as
A and B or C	(A and B) or C
A or B and C	A or (B and C)
(A or B) and C	(A or B) and C
A or B and C and not D	(A or (B and (C and not D)))
C and not D and A or B	(((C and not D) and A) or B)
((A or B) and C) and not D	((A or B) and C) and not D
C and not D and A near3 B	(C and not D) and ((A near3 B))

Source: Infomine Search Tips (April 2006)

Three sample queries to find resources relevant to "OAI-PMH," "metadata," and "access within four words of knowledge," show the wide variation in results retrieved from INFOMINE and other general metasearch and cross-archive search engines. Overall,

---

[196] *NSDL Whiteboard* report, issue 61, October 2004, available from
http://content.nsdl.org/wbr/Issue.php?issue=61.

INFOMINE's results are the narrowest (or most refined) but in some instances, such as the query to find resources relevant to OAI-PMH, they appear too limited. INFOMINE's coverage of metadata is stronger and it is the only search engine to support proximity operators. However, it shares the unsolved problem of duplicate entries with OAIster (and probably with the other services as well). This primitive exercise demonstrates the need for a more thorough study to better understand the strengths of these service entities (e.g., Scirus's coverage of journal articles). It also underscores the reason why users need a thorough understanding of the universe covered by these search engines (cum databases) and the need for "nutrition and ingredient labeling" as discussed elsewhere in this report and proposed by Péter Jascó in 1993.

Table 34: Comparative Search Results: INFOMINE, RDN, OAIster, Scirus & Google Scholar

<OAI-PMH>	<Metadata>	<access near4 knowledge>
INFOMINE • 1 expert-selected record (Emory University's MetaScholar Initiative) • 10 robot-selected records (including articles from *D-Lib* and *Ariadne*)	**INFOMINE** • 230 expert-selected records • 584 robot-selected records	**INFOMINE** • 15 expert-selected records • 78 robot-selected records including many duplicates from Knowledge Management Think Archive and MayoClinic.com
RDN • 3 results (Grainger Engineering Cross-Archive search service, OAIster, and Project Euclid)	**RDN** • 147 results	**RDN** • does not support proximity searching, phrase control or AND operator • <access knowledge> returns 468 entries • <access OR knowledge> returns 22,961
OAIster • 181 items including noticeable duplication	**OAIster** • 143,392 items • >128,700 from 3 institutions w/ "metadata" in name	**OAIster** • does not support proximity searching • <access AND knowledge> retrieves 6,339
SCIRUS • 9 journal results • 2,235 preferred Web results; when Hong Kong U of Science & Technology (HKUST) is excluded, results are reduced to 77 items • 10,977 other Web results; without HKUST results are reduced to 6,766	**SCIRUS** • 2,263 journal results • 16,729 preferred Web results • 962,865 other Web results	**SCIRUS** • does not support proximity searching • <access AND knowledge> returns: • 107,589 journals results • 69,138 preferred Web results • 3,359,813 other Web results
GOOGLE SCHOLAR	**GOOGLE**	**GOOGLE SCHOLAR**

• 1,420 items	**SCHOLAR** • 266,000 items

• 1,420 items	**SCHOLAR** • 266,000 items	• does not support proximity searching • ALLTHEWORDS: <access knowledge> returns 1,830,000 results

4.5.3 Intute (formerly RDN—Resource Discovery Network)

Update Table 31: Intute based on DLF Survey responses, Fall 2005

	Intute (as of mid-2006) http://www.intute.ac.uk/ **Resource Discovery Network (RDN)** http://www.rdn.ac.uk/ (former name)
ORGANIZATIONAL MODEL	Partnerships: 8 universities as host institutions and more than 70 collaborators (educational and research organizations).
SUBJECT	Multidisciplinary. Covers: Arts and humanities; social sciences; science, engineering, technology; and geography; health and life sciences.
FUNCTION	To advance education and research by promoting the best of the Web through evaluation and collaboration. RDN's vision is to create knowledge from Internet resources and in doing so, enable people to fulfill their potential. RDN brings together the best Web sites for education and develops associated services to embed these resources in teaching, learning and research.
PRIMARY AUDIENCE	Academic community
STATUS	As of late 2005, RDN comprises of eight subject hubs. After a period of review, analysis and internal consultation, RDN aims to build upon and re-establish its position in the further and higher education environment and in the Internet information environment. To this end RDN will: • Move to a new organizational structure • Integrate hardware and software platforms • Introduce a more holistic performance measurement framework • Implement the outcomes of a strategic branding exercise and review of visual identity so an established service will move into a new mode of delivery in mid-2006, and into a third phase of evolution.
SIZE	Number of records (as of July 2005): Altis 4,020; artifact 5,500; BIOME 30,700; EEVL 12,415; GEsource 8,400; HUMBUL 10,000; PSIgate 13,500; SOSIG 26,800. TOTAL: 111,335
USE	Per month: ~12 million pages served; ~740K Internet (RDN catalog) searches
ACCOMPLISHMENTS	1. Launch of the GEsource World Guide service. 2. Additions and updates to the Virtual Training Suite of online

	Internet tutorials. 3. Creation of the RDN Executive at MIMAS, Manchester Computing, the University of Manchester and the start of the strategic change process for the RDN.
CHALLENGES	1. Differentiating the service from search engines. 2. Embedding the service or parts of the service in VLEs and more widely in the learning, teaching and research process. 3. Understanding changing user needs in terms of subject searching, indexing, level of description of resource required.
TOOLS OR RESOURCES NEEDED	1. Automatic metadata creations tools. 2. Visualization technology. 3. Text mining tools. 4. Cross-searching technologies. 5. Portal technology
GOALS OF NEXT GENERATION RESOURCE	The name of the next generation of the RDN will be Intute. The next generation Web site aims: - To establish Intute as the successor to the RDN and its Hubs, where existing users can find the services they know. - To attract and retain new users of Intute. - To differentiate Intute from search engines and gateways. - To convince students, researchers, academics, teachers, and librarians / intermediaries to use Intute to make intelligent use of the Internet for education and research. - To promote the people who create Intute and convey the concept of the service as authoritative mentor of the Internet.

In mid-2006 after re-structuring and re-branding, the former "RDN" débuted as "Intute," a new name combining Internet and Tutorial to connote the amalgamation of guided learning and online resource discovery.[197] As was the case with RDN, Intute is a free online service of high-quality Web resources for education and research, selected by a network of subject specialists. The new service consolidates RDN's eight subject gateways into four broad subject areas, bringing them together with a unified interface:

- Arts and Humanities
    - Artifact: Arts and Creative Industries
    - Humbul: Humanities
- Health and Life Sciences
    - BIOME
- Science, Engineering and Technology
    - EEVL: Engineering, Mathematics and Computing
    - GEsource: Geography and the Environment
    - PSIgate: Physical Sciences
- Social Sciences
    - Altis: Hospitality, Sports, Leisure and Tourism
    - SOSIG: Social Science Information Gateway

---

[197] Information cited from Intute development and FAQ: http://www.intute.ac.uk/development/

Intute aims to support interdisciplinary inquiry while still providing the same level of subject access via its domain-specific Internet resource catalogs. It serves as a resource base for integration into a variety of e-learning platforms (or VLEs—virtual learning environments) and discipline-specific portals, as evident by the incorporation of EEVL into the pilot engineering metasearch service, PerX (discussed in section 4.2.8 of this report).[198] Intute will be developed so it can be integrated more easily into institutional portals (and VLEs) whereby its resources/contents may re-emerge in customized subject-based portals, created according to local needs. The Intute Virtual Training Suite provides subject-based e-learning tutorials and resurrects the general training sequence intended to teach critical thinking skills, known as the Internet Detective.

Intute is a core service of JISC hosted by MIMAS (Manchester Information & Associated Services at the University of Manchester). Its operations adhere to policies and standards documented through a formal Service Level Agreement that covers: collection management policy; marketing and communications; strategic plan; technical integration plan; cataloging guidelines; and network services such as format conversion, printing, authentication and e-commerce.[199] JISC monitors and audits Intute's performance and produces quarterly service "trend data" consisting of statistical graphs charting the number of catalog searches, Web pages served, and HelpDesk inquiries. As of late April 2006, the data about RDN and its constituent services hubs is current to October 2005.[200] In the first of a two-part series in *Ariadne*, Hiom (2006) provides a "Retrospective on the RDN," along with a timeline of its milestones (http://www.rdn.ac.uk/projects/eprints-uk/). A future article will discuss the strategies underlying its transformation into Intute.[201]

---

[198] PerX is available from http://www.icbl.hw.ac.uk/perx/. See the Project Deliverables for future plans to embed PerX in VLEs and other strategies.
The ePrints UK Project is available from http://www.rdn.ac.uk/projects/eprints-uk/.
[199] See Annex A: Services Provided by RDN. RDN Service Level Definitions, 1st August 2005 to 31st July 2006. Available from http://www.mu.jisc.ac.uk/slas/rdn/rdnsld2005-06.html.
[200] RDN service trend data is available from http://wiki.library.oregonstate.edu/confluence/display/DLSRW/WorkshopResources.
[201] Intute had not yet launched when this report was submitted so the author was unable to test out its functionality.

4.5.4 California Digital Library (CDL) Metasearch Initiative

Update Table 32: CDL Metasearch Infrastructure Project based on DLF Survey responses, Fall 2005

	CDL Metasearch Infrastructure Project http://www.cdlib.org/inside/projects/metasearch/
ORGANIZATIONAL MODEL	California Digital Library and UC campus libraries, partially grant-funded.
SUBJECT	Specific to each instance created via the infrastructure.
FUNCTION	Build localized metasearch services tailored to a particular audience and/or need.
PRIMARY AUDIENCE	Academic community
STATUS	Under development
SIZE	The first instance will include harvested metadata from 35,000 OAI records, along with at least 5 licensed databases.
USE	Not yet available
ACCOMPLISHMENTS	1. Significant progress in integrating vendor metasearch product (ExLibris's MetaLib) with CDL's Common Framework software infrastructure. 2. Creation of a prototype harvesting tool (based on OAIster), harvesting, and evaluation of both harvest and harvesting tool. 3. Establishment of an SRU-compliant gateway to OAI harvested metadata.
CHALLENGES	1. Gaining a better, more specific understanding of user needs (and how needs may vary depending on the institution and/or type of user). 2. Translating the prototype(s) into a production service. 3. Supporting the service once it is in use.
TOOLS OR RESOURCES NEEDED	Robust, flexible, open source tools for metadata normalization and enrichment, Web crawling, indexing, and searching, and widespread implementation of protocols (e.g., SRU) by vendors.
GOALS OF NEXT GENERATION RESOURCE	To enable the easy discovery of appropriate metasearch portals, or even to dynamically select the resources to be metasearched at the moment of query.

The California Digital Library (CDL) Metasearch Infrastructure Project aims to leverage CDL's experience over the past six years since it first deployed "SearchLight," by establishing an infrastructure that will enable UC campus libraries to create customized search portals for specific audiences and purposes. The metasearch infrastructure adheres to the principles and standards set forth in CDL's Common Framework, an open, services-oriented technical architecture that provides an integrating framework for a full-spectrum of library services, ranging from archival (where objects are stored

locally, e.g., UC's Digital Preservation Repository), to metadata only (where only metadata is stored locally), to portals (where no data is stored locally).[202]

In addition to The American West metadata portal, discussed earlier in this report, the CDL has several other prototype portals in various stages of development: [203]

- NSDL: In fulfillment of a NSF grant to build and enhance the NSDL, the Earth Sciences portal is geared to meet the needs of the UC geosciences community, and serve as an exemplar of integrating NSDL content into university library services. A pilot deployment of the Earth Sciences portal is being evaluated as of mid-May 2006.

- SmartStart: Targeted to meet the needs of undergraduates and others outside of their area of expertise.

- Discipline-Specific: The first deployment is targeted to meet the needs of with faculty and graduate students in European studies (Western, Central and Eastern Europe, including Russia).[204]

An important background document, "Integrating Information Resources: Principles, Technologies, and Approaches" (Christenson and Tennant 2005) summarizes findings from CDL's studies of user needs relative to integrated searches. They report that from a user perspective, metasearch tools must exhibit:

- *Speed and simplicity* of the Internet search engines (Google).

- *Convenience* of e-commerce (Amazon). Participants' Internet usage has set high expectations for a service-rich environment.

- *Reliability, authority and integrity* of information resources that are trusted because of the brand they carry (whether imparted by a prestigious library, academic institution, professional society, or even a state education curriculum. (Christenson and Tennant 2005, 3)

The report also fleshes out the content discovery and integration principles that should inform the design of CDL's metasearch services.

---

[202] For more information about CDL's Common Framework refer to http://www.cdlib.org/inside/projects/common_framework/index.html.

[203] See Projects in progress at http://www.cdlib.org/inside/projects/metasearch/portals.html.

[204] Refer to UCLA European Integration Report: Metasearch Assessment (June 2005). Available from http://www.cdlib.org/inside/assess/evaluation_activities/docs/2005/uclaMetasearchReport_june2005.pdf.

Table 35: CDL's Metasearch Infrastructure Principles

Content Discovery Principles	Integration Principles
1. *Only librarians like to search, everyone else likes to find.*[205].	1. *Integrate metadata whenever possible.*
2. *"Good enough" is just that.*	2. *Exploit metadata similarities.*
3. *All things being equal, one place to search is better than two or more.*	3. *Honor metadata differences.*
4. *What is not searched is as important as what is.*	4. *Offer appropriate methods to narrow the scope.*
5. *Place services as close to the user as possible.*	5. *If you can't centralize metadata, centralize searching.*

Source: Christenson and Tennant 2005, 4-6.

The authors then chart the strengths of five different methods of integration (e.g., ingesting, harvesting, Web crawling, syndicating, and metasearching) against the relevant integration principles, the conditions in which each method is the most appropriate, and the implementation obstacles.

Table 36: Metasearch Integration Methods and Practices

	Enable Content Submission (Ingest)	Harvest Metadata (OAI-PMH)	Crawl Web Sites	Enable Content Syndication (RSS)	Enable Federated Queries (Metasearch)
Relevant integration principle(s)	• All appropriate metadata stored internally in a common format uniformly applied	• All appropriate metadata stored internally in a common format uniformly applied • Honor metadata differences • Offer appropriate methods to narrow the scope	• Integrate metadata whenever possible	• Integrate metadata whenever possible	• If you can't centralize metadata, centralize searching
When is this method appropriate?	• Local collection that will be locally accessed • Content is relatively stable • Resources available to provide rich native interface	• Need access to large collections you don't want to have in-house • Need a fast search	• To provide search access to a targeted collection of web sites	• Provide access to frequently updated content or news – current awareness	• When metadata cannot be centralized • When it is too time consuming for users to access multiple resources separately • Resource discovery • When users will need to find "just a few good things" • When content is frequently updated

---

[205] For more information refer to Roy Tennant's Webcast, "Metasearching: Librarians like searching, users like finding" (February 8, 2005), available from http://infopeople.org/training/webcasts/02-08-05_metasearch.html.

| What are the obstacles? | • May not want to have "ownership" responsibilities
• Storage space (at a very large scale) | Mostly obstacles related to providing access:
• Normalization of metadata
• Duplication of records –aggregate providers
• Varying levels of granularity amongst digital objects
• Contextualizing results

And
• Accounting for XML validation errors | Mostly obstacles related to providing access:
• How should search results be presented? By individual web page? By web site, then by page? | • At this point in time, still a limited number of resources in this format

• Range of options yet to be fully explored | • Lack of standards
• Avoiding "lowest common denominator" interface – losing benefits of native interface(s)
• Staff training
• Maintenance time/costs
• De-duping difficulties and vendor concerns about duplicate display
• Vendor concerns about server overload (as target)
• Contextualizing results
• Inadequate or non-existent search result ranking |

Source: Reprinted with permission of the authors. (Christenson and Tennant 2005, 7)

As the authors explain:

> A suitably developed metasearching infrastructure can be used to provide a common interface to content integrated by any or all of these methods. Thus the standard metasearch application marketed by software vendors is but one piece of a robust metasearching infrastructure. Such an infrastructure must be capable of using each of the integration techniques identified in the above chart while providing a unified user interface to the whole. (Ibid, 7)

The schematic bellows depicts how users would access digital resources via different portals. They would also have a suite of tools readily available to manage citations and facilitate the re-use, manipulation, annotation, and integration of resources into teaching and research platforms (e.g., the Scholars Box).[206] Current under development "the Scholar's Box is a tool that gives users "gather/create/share" functionality, enabling them to gather resources from multiple digital repositories in order to create personal and themed collections and other reusable materials that can be shared with others for teaching and research. The Scholar's Box can currently perform the following functions:

---

[206] The Scholars Box, under development as part of UC Berkeley's Interactive University Project, *is conceived and designed to address important interoperability issues at the intersection of four key information technology domains: digital libraries and repositories; educational technologies and learning management systems; web syndication and portal technologies; and desktop applications and structured content authoring tools.* Refer to: http://interactiveu.berkeley.edu:8000/IU/July2003News#sb.

- **Gather**: From California Digital Library, amazon.com, google.com, NSDL, RSS feeds, METS (digital library), WWW, CDL's metasearch system, and the local file system.
- **Create**: Data and metadata gathered, annotated, and organized into personal collections via drag and drop
- **Share**: IMS-CP, OpenOffice.org Presentation or Text document, PDF, HTML, a METS document, a set of Endnote references, Chandler Parcel, or sent to a weblog via the Blogger API" (Raymond Yee[207]).

The CDL infrastructure relies on a combination of open-source and commercial solutions. For example, CDL chose Ex Libris's MetaLib to enable access to commercial databases, externally-managed resources, and the Melvyl online catalog.[208] MetaLib interoperates with the Metasearching Infrastructure that manages other components in the context of CDL's Common Framework.

---

[207] Raymond Yee, Scholars Box wiki: http://raymondyee.net/wiki/ScholarsBox.
[208] Information about Ex Libris's MetaLib is available from http://www.exlibrisgroup.com/metalib.htm. It is possible to register to listen to the 60-minute archived "Webinar" about this product.

Figure 45: CDL Metasearching Infrastructure Schematic
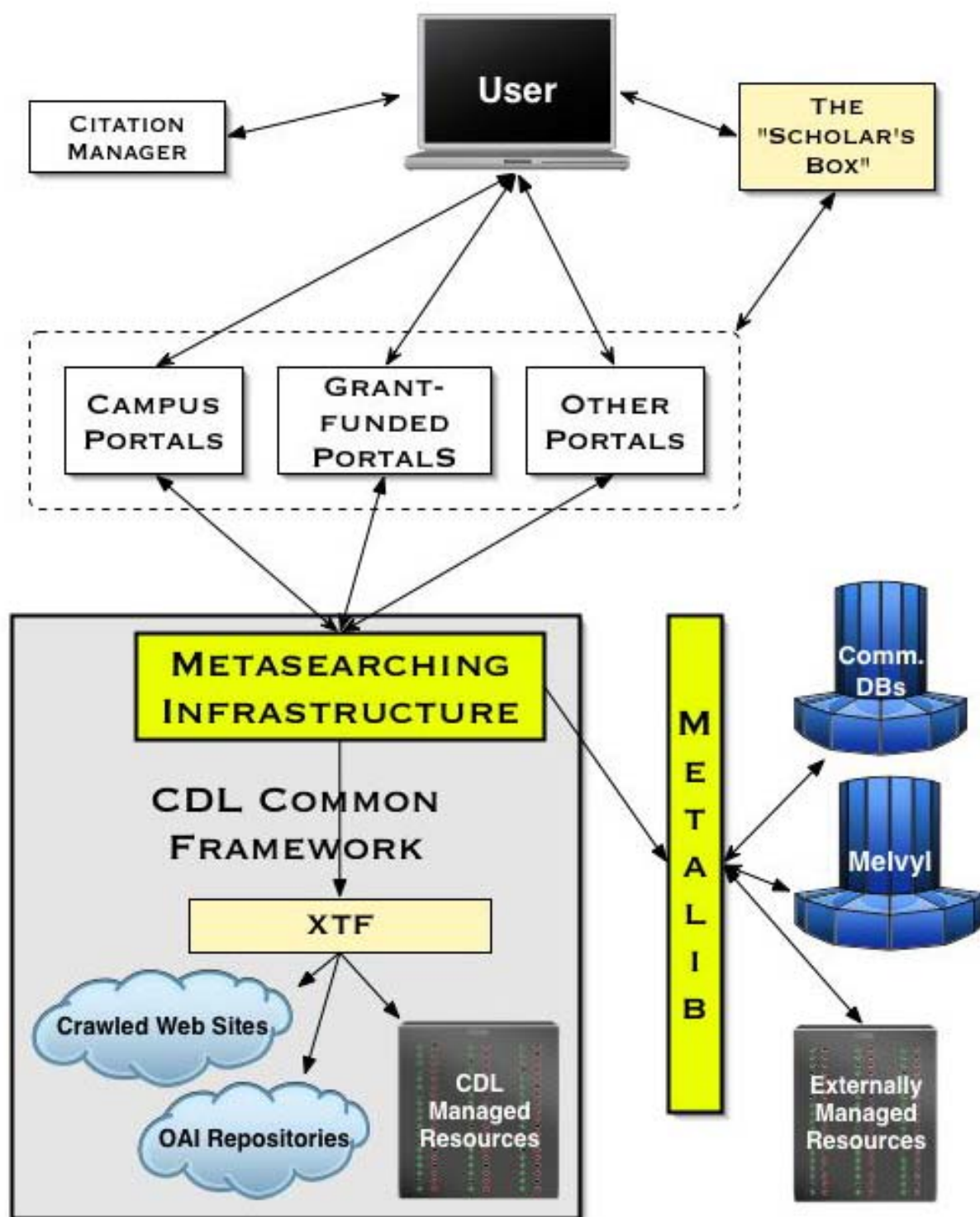
Source: http://www.cdlib.org/inside/projects/metasearch/diagramCF.jpg . Used with CDL's permission.

4.5.5 Current Issues and Future Directions

The deployment of these services addresses some of the issues discussed earlier in this report in response to the "Amazoogle Effect," by providing users with the types of services they have come to expect. These efforts are abetted by the work underway at NISO to develop standards and specifications that enable search and retrieval across multiple platforms and vendors, and linking to "appropriate" resources through OpenURL resolution systems. This work is carried out by NISO's OpenURL Framework for Context-Sensitive Services (http://www.niso.org/committees/committee_ax.html) and the NISO Metasearch Initiative. The second initiative brings together three major stakeholder groups organized into three cross-sector task groups dealing with Access Management; Collection and Service Descriptions; and Search and Retrieval specifications (Hodgson, Pace and Walker 2006). The overall goal is "to move toward industry solutions NISO sponsored a Metasearch Initiative to enable:

- metasearch **service providers** to offer more effective and responsive services;
- **content providers** to deliver enhanced content and protect their intellectual property; and
- **libraries** to deliver services that distinguish their services from Google and other free web services. http://www.niso.org/committees/MS_initiative.html)

Available as of July 2005, the NISO Metasearch XML Gateway (MXG) Implementers' Guide (version 0.3) describes the MXG protocol that enables service providers to expose their content and services to a metasearch engine. (Such a gateway has been implemented, for example, by Berkeley Electronic Press's ResearchNow portal, described in section 4.2.11).

The first set of deliverables and recommendations was presented at a NISO workshop in September 2005; these documents are available along with the workshop presentations at the NISO Metasearch Initiative Web site. Among the important recent developments, the NISO Z39.92-200x, Information Retrieval Service Description Specification, was released by the Collection and Service Descriptions for trial use through October 2006. "This standard defines a method of describing Information Retrieval oriented electronic services, including but not limited to those services made available via the Z39.50, SRU/SRW, and OAI protocols. The ZeeRex standard addresses the need for machine readable descriptions of services in order to enable automatic discovery of and interaction with previously unknown systems. It specifies an abstract model for service description and a binding to XML for interchange."[209]

---

[209] Available from http://www.niso.org/standards/resources/Z39-92-DSFTU.pdf.

Library service vendors, as active contributors to and beneficiaries of the NISO Metasearch Initiative, are entering the metasearch market and designing new applications based on layered architectures that are intended to consolidate information search results and meet user needs from "discovery to delivery," as exemplified by Ex Libris's new "Primo" metasearch architecture below.
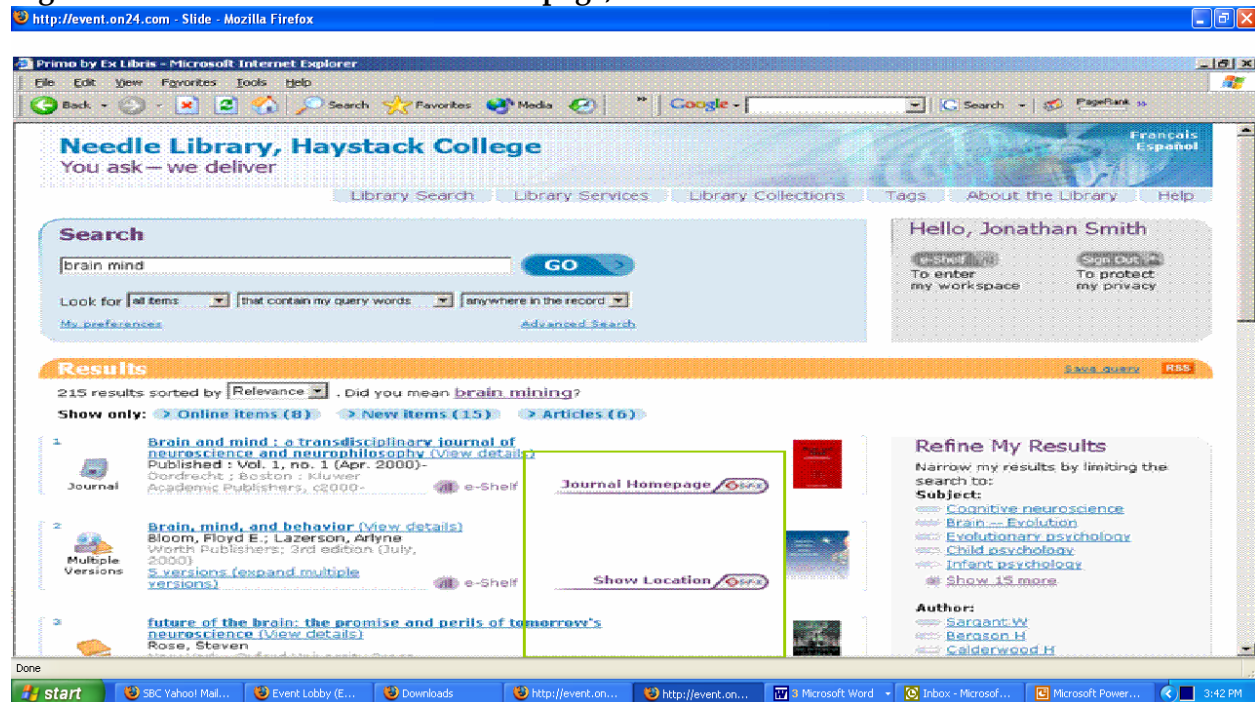
Figure 46: Screenshot of Primo Architecture, Ex Libris
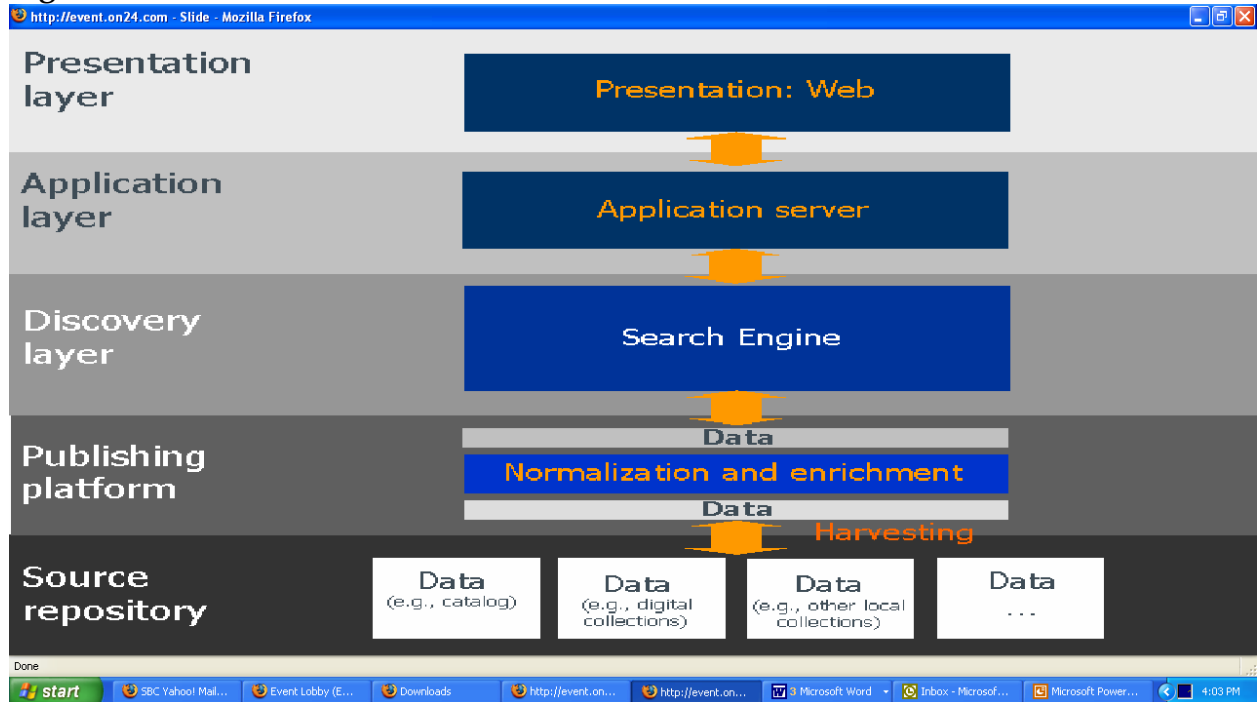
Source: Webinar presentation, "Primo: an Exclusive Peek from Ex Libris," Tamara Sadeh, May 9, 2006. Reproduced with permission.

This architecture helps to "create a superb user experience layer, decoupled from back-office functions, separating data creation and maintenance from its discovery." The publishing platform enables libraries to leverage resources irrespective of source, enrich the data, and expose hidden collections. Meanwhile the user is presented with a system that recognizes him (Hello, John Smith) and with results that can be refined, extended, altered, and displayed in various ways, as exemplified by the two prototype screenshots below (Tamar Sadeh, Ex Libris Webinar, May 9, 2006).

Figure 47: Screenshot of PRIMO search page, Ex Libris

Source: Webinar presentation, "Primo: an Exclusive Peek from Ex Libris," Tamara Sadeh, May 9, 2006. Reproduced with permission.

Figure 48: Screenshot of PRIMO search options
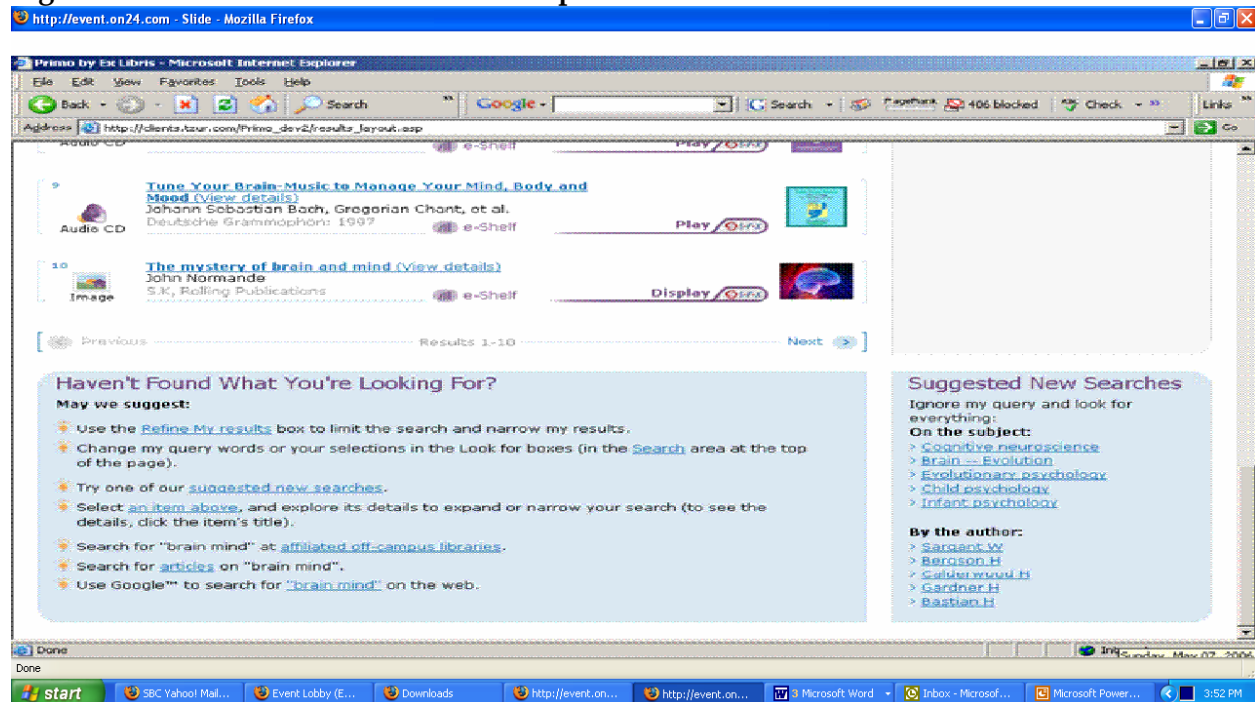
Source: Webinar presentation, "Primo: an Exclusive Peek from Ex Libris," Tamara Sadeh, May 9, 2006. Reproduced with permission.

The California Digital Library is uniquely positioned to develop its own system wide metasearch infrastructure, relying on a combination of locally developed, open source and proprietary tools and systems. A challenge for all academic libraries is evaluating the appropriate balance of components and services developed internally versus those they purchase externally. With more than 600 people from 29 countries participating in Ex Libris's early preview of PRIMO, it seems that many libraries have already begun to consider their options.

Finally, it is worth reiterating that metasearch goal in this context is not about simplified "one-stop-shopping," but about creating a distributed information environment that can deliver subsets of resources, services and tools to users according to their particular needs.[210]

---

[210] Refer to Lorcan Dempsey's Weblog, "From Metasearch to Distributed Information Environments," October 9, 2005 where he reports on the fall 2004 NISO OpenURL and Metasearch meeting: http://orweblog.oclc.org/archives/000827.html

5.0 Conclusions

This section compares services' baseline features in 2003 and 2006 (i.e. organizational model, subject, function, primary audience, status, size, and use). Next, problems encountered while preparing this report are enumerated in "an embarrassment of glitches," highlighting the need for better product and service "nutrition and ingredient" labels (Jacsó 1993). Progress towards addressing three primary issues and five future directions from the 2003 report is discussed before giving attention to "the pulse" in 2006. Growth in adoption of OAI, coupled with a better understanding of its potential uses and limitations; interoperability as an international phenomenon; sustainability and funding; and next generation service characteristics are highlighted. The report closes with a summary of ten "imperatives" for successful services.

5.1 Comparison of 2003 and 2006 Baseline Features

5.1.1 Organizational Model

2003
• Integral to issues of quality assurance, economic viability and long-term sustainability.
• Almost all sites under review are sponsored by institutions of higher education or governmental agencies.
• Many are promoted by a handful of key individuals.
• Few are fully integrated into a broad-based organizational structure.
• Many address R&D issues and have not transitioned to full production.
• Almost none have a business plan.
• Some rely on community-based input and collaboration with varying degrees of formal governance structures.
• Most are developed with external support.

2006 Survey Responses and Observations

- Although relatively few respondents noted organizational changes in the survey, upon closer examination there were numerous shifts in administrative and governance structures, especially to anchor the service more securely to the operation of an established institution or disciplinary group.
- The services under review are still predominantly sponsored by institutions of higher education or governmental agencies, but there is increased connectivity with disciplinary organizations.
- As these services mature, they are becoming more fully integrated into established institutions and a widening circle of collaborators assume advocacy roles.

- A few services are beginning to turn to libraries to fulfill implementation and preservation functions while they continue in R&D, prototyping services.
- More services are developing business, marketing, and fiscal sustainability plans. Typically, these are hybrid approaches, including a mix of institutional, grant/foundation, and revenue-producing streams. Nevertheless, there is still widespread concern about future funding and long-term viability.
- Community-input is viewed as an essential ingredient in developing most services. A great deal of emphasis is placed on creating active communities of practice.
- When implementing service-oriented architectures (SOA) industry analysts advise that

  > *building a governance framework is a critical early milestone on the road to a successful SOA implementation – not a governance framework for the SOA implementation specifically, but rather, a framework that outlines governance best practices across the organization that will leverage the power and flexibility of the Services that form the core of the SOA implementation.* (Bloomberg 2006)

- The UK and Australian e-Framework for Education and Research reflects this approach in so far as they began the process by adopting principles to guide the partnership (http://www.e-framework.org/about).

5.1.2 Subject Coverage

2003
- Major initiatives cluster around funding agencies in the sciences and cultural heritage.
- Communities of practice formed around disciplines; audiences; type of media; software; or philosophy.
- Published literature on disciplinary differences in scholarly communication appears primarily in the sciences.
- Much of the literature produced by PIs.
- Some mainstream news coverage focusing on the economic dynamics of the open access movement.

2006 Survey Responses and Observations

- The NSF, IMLS and The Andrew W. Mellon Foundation have a tremendous influence in supporting the development of the services under review in this report. Although the predominant focus here is on the sciences and cultural heritage, in fact there is increasing activity across a full spectrum of disciplines and subject areas. Despite few social science examples in this report, there are significant activities underway as evident from such activities as Cyberinfrastructure initiatives, the leading role of the Inter-university Consortium of Social Science Research in the archiving of digital datasets, new affiliations between the American Economics Association and RePEc, the development of Nereus in Europe and so forth.

- Communities of practice continue to form around disciplines, audiences, types of media, technology platform, and philosophy. To this list one must add communities focusing on e-learning, e-research, Web publishing, digital preservation, and e-administration (including records management). While services may be aligned primarily with one community, it is increasingly apparent that they need an understanding of—if not direct engagement with—multiple communities in order to garner the requisite combination of subject, technology, and service-environment expertise.
- The literature on scholarly communication now crosses all disciplines.
- To meet their responsibilities as researchers, PIs continue to contribute substantially to the literature, but their efforts are now joined by a widening circle of authors, extending from practitioners and journalists to researchers and theoreticians.
- There is phenomenal growth in media coverage of issues under review in this report. Open access, scholarly information practices, mass digitization, the digital divide, publicly-funded research, copyright and fair use are all part of the public discourse.

5.1.3 Function

2003
- Conflicting and overlapping definitions of concepts (e.g., digital libraries, portals).
- Service are complex and do not lend themselves to solitary functional "encapsulation."
- Dynamic and innovative nature of these services fuels their capacity to change functionality or scope.
- Successful data providers attract multiple new services, creating new levels of aggregation and customized functionality.

2006 Survey Responses and Observations

- Very few of the services under review changed their core function since 2003, although there are some elaborations and modifications in scope. Intute (formerly RDN-Resource Discovery Network), for example, notes: *We exist to advance education and research by promoting the best of the Web through evaluation and collaboration. Our vision is to create knowledge from internet resources and in doing so, enable people to fulfill their potential. We bring together the best websites for education and develop associated services to embed these resources in teaching, learning and research.* Other new initiatives represent a variety of functional models ranging from DLF Aquifer's service-oriented approach to SouthComb's portal development.
- Data providers continue to morph into service providers (dLIST spawns DL-Harvest) and vice versa (OAIster makes it metadata available to other services).
- Blinco and McLean's "Wheel of Fortune" (section 2.2.1) depicts the different communities of practice and dimensions of the scholarly information environment, including Web publishing, e-learning, e-research, administrative computing and scholarly information.

- Discussion continues unabated about how best to differentiate among concepts such as repositories, archives, digital libraries and portals; however, there are several sustained efforts to define their distinctive qualities. Heery and Anderson (2005) distinguish digital repositories from other digital collections according to four characteristics:
    - content is deposited in a repository, whether by the content creator, owner or third party
    - the repository architecture manages content as well as metadata
    - the repository offers a minimum set of basic services e.g., put, get, search, access control
    - the repository must be sustainable and trusted, well-supported and well-managed
      (p. 2)

  They then develop a typology of repositories by content type, coverage, functionality and target user group. Among primary functions, Heery and Anderson propose:
    - Enhanced access to resources (resource discovery and location)
    - Subject access to resources (resource discovery and location)
    - Preservation of digital resources
    - New modes of dissemination (new modes of publication)
    - Institutional asset management
    - Sharing and re-use of resources [e.g., datasets and learning objects] (p. 14)

  As discussed earlier in this report, this typology formed the basis of the disciplinary landscape analysis for engineering and ensuing cross-archive search service, PerX (http://www.engineering.ac.uk/).

- JISC provides definitions of major service components in the context of its Information Environment Architecture (Powell 2005). For example:
    - Portal: A network service that provides a personalised, single point of access to a range of heterogeneous network services, local and remote, structured and unstructured. Portal functionality often includes resource discovery, email access and online discussion fora. Portals are intended for (human) end-users using common Web 'standards' such as HTTP, HTML, Java and JavaScript. In the context of the JISC IE, portals interact with brokers, aggregators, indexes, catalogues and content providers using Z39.50, SRW, the OAI-PMH and RSS/HTTP.
    - Aggregator: A structured network service that gathers metadata from a range of other, heterogeneous, local or remote structured network services. Aggregators are intended for use by software applications. In the context of the JISC IE, aggregators interact with indexes, catalogues, content providers and other aggregators using the OAI-PMH and RSS/HTTP. Aggregators interact with portals using the OAI-PMH. In some cases an aggregator may offer its aggregated metadata as a Z39.50 target.
    - Subject Gateway / Gateway: A network service based on a catalogue of Internet resources. The gateways provided by RDN [now Intute] hubs focus

on particular subject areas. (JISC Information Environment Architecture, Glossary, http://www.ukoln.ac.uk/distributed-systems/jisc-ie/arch/glossary/)

5.1.4 Audience

> **2003**
> - Counter prevailing trends: serving multiple audiences for different uses versus serving specialized audience for restricted uses.

2006 Survey Responses and Observations

- Several services comment on the difficulties of attempting to meet a wide range of audience needs and expectations. This is particularly true for deployments attempting to serve the spectrum of K-12 and higher education clienteles. As more user studies identify differences in the work environments, habits, and traditions of instructors by grade-level and discipline, broad-based services struggle to effectively tailor subsets of resources and tools for targeted use. Theoretically, service-oriented architectures are designed with the flexibility of meeting this challenge and portal toolkits (e.g., NSDL's Scout Portal Toolkit and Collection Workflow Integration System, see Almasy 2005) are intended to facilitate customization.
- More hybrid services, offering a combination of open and restricted use, are appearing. As commercial journal publishers enter the "open choice" market so too are OA service providers starting to integrate restricted use resources.
- As services attempt to sustain themselves without benefit of grant funding, they are instituting different level of access, with value-added services and benefits to members or subscribers.

5.1.5 Status

> **2003**
> - Status is a moving target; most services are characterized as "evolving."

2006 Survey Responses and Observations

- Most services now consider themselves "established."
- Several new efforts are clearly pilots. While their future is uncertain, these undertakings serve as building blocks for more durable systems.
- A number of services are suspended in perpetual beta.

5.1.6 Size

> **2003**
> - Difficult to measure and interpret.
> - Can change rapidly.
> - A limited number of archives may account for the majority of records.
> - Paradox of size: critical mass is important but may also inhibit customization for specific uses.

2006 Survey Responses and Observations

- At the individual service-level, statistics about size and what is measured (number of metadata records, full-object links, full-text articles, collections, repositories, free and restricted use resources, etc.) is still difficult to obtain.
- All services increased in size; many noted growth as one of their major accomplishments. Overall size continues to change rapidly although for some sectors (e.g., full-text or peer-reviewed items; IR deposits) growth is incremental.
- As discussed below and evident from the prior review of "OAI demographics," there are more tools available to obtain a composite picture of OAI growth and distribution.
- As discussed above, the critical mass versus customization dialectic remains a challenge.

5.1.7 Use

2006 Survey Responses and Observations (usage data not collected in 2003)

- Usage statistics are even more problematical to obtain and interpret. Numerous services do not make their usage data readily available. A few services, not surprising primarily those aggregating self-archived research papers, are exemplary in making their usage data transparent, e.g. arXiv, http://www.arxiv.org/todays_stats; dLIST, http://dlist.sir.arizona.edu/es/index.php?action=show_detail_date;range=4w.
- Growing interest in Webmetrics and efforts to incorporate OAI sources (as described in section 4.2.10) should help to bring more consensus, if not standardization, in usage measures.

5.2 An Embarrassment of Glitches

Preparing this report and testing various services was not without its frustrations. Among the glitches encountered:

- No information or misinformation about the scope and attributes of resources harvested
- Moribund harvests resulting in stale crops
- No information or misinformation about frequency of harvests
- Service disruptions of several weeks to several months duration without any indication to the user
- Perpetual Beta—enduring from 2003 to present
- Advanced feature malfunctions
- Programming "bugs" grossly affecting search results
- Duplicate harvests—the same data provider aggregated twice by the service provider
- Broken links leading nowhere
- Links leading to restricted resources without any indication to user
- Duplication of items within a service
- Out-of-date collection/repository descriptions
- Out-of-date wikis and empty templates where current news is anticipated
- Most recent "news" is a year or more out-of-date
- Widely varying resource and usage statistics provided within the service
- Lack of internal agreement about what is measured and how to measure it

In a 1993 guest editorial appearing in "Database," Péter Jacsó harkens back to Jeff Pemberton's decade-old plea for "exposing the problems of dirty data." Noting that the situation had only grown worse in the ensuing years, Jacsó takes the suggestion a step further, proposing "nutrition and ingredient" labels for databases. Now, more than 20 years have past since the original article—and the need is transferred to the new scholarly information environment on the Web.

An adaptation of Jacsó's database nutrition and ingredient label could serve as a starting point, regularly reporting such items as: number of records; quarterly increase in size; time-lag since last update (proportion of database that is current); record of service availability (in last week, month, quarter, year); source coverage (depth, breadth, geographic provenance); content (types of materials, languages, subjects, restricted use versus freely available, full-object versus bibliographic metadata only); access points (percent of records that include major search fields, e.g., title, author, subject, publication year), and "transfat" (estimated percentage of duplicate records).

"Centers of value" formulated in conjunction with the review of faculty needs in using digital resources provides a useful "product-level" summary (elements that might be elaborated on, for example, in a collection development policy):

Table 37: Digital Resources and Centers of Value

- Content coverage (chronological, geographic, thematic, disciplinary, type of "original"—manuscripts, coins, maps, games)
- Form of representation (i.e. availability of digital formats and portability, e.g., jpeg, tiff, sid; proprietary or open, level of metadata: structured, standard, rich or thin; wrapper issues, e.g. HTML, XML,METS)
- Authority (e.g., source, maintenance, institutional affiliation)
- Permitted uses and digital rights of reuse
- Persistence (e.g., how long is the resource up, how often does updating occur?)
- Exposure for discovery (e.g., searching paths, browsing, availability for federated search, availability for Google crawling)

Source: Harley et al. 2006, 41; based on suggestion of Arnold Arcolio, RLG.

5.3 Updates: 2003 Issues and Future Directions

The 2003 report identified three critical issues:

1. The absence of a user-friendly comprehensive registry of OAI-compliant services geared towards users to improve resource discoverability.
2. The lack of priority given to creating and exposing OAI-compliant metadata to meet minimal let alone enhanced standards, coupled with problematic issues of granularity and the need to amass more object-level data.
3. The aggregations did not provide users with a meaningful "context" or match the level of refinement available from the resource's native environment or of their proprietary counterparts.

It concluded by highlighting five future directions to pursue: (1) giving more attention to users and uses; (2) finding solutions to digital rights management and digital content preservation; (3) building personal libraries and collaborative workspaces; (4) putting digital libraries in the classroom and digital objects in the curriculum; and (5) promoting excellence.

These issues and directions are updated below. Accomplishments and challenges regarding shareable metadata are more fully discussed earlier in this report (section 3.1.3).

5.3.1 Registries, Metadata, and Placing Objects in Context

Considerable progress is evident in addressing the three concerns specified in the 2003 report. First, various new or enhanced registries, directories, and tools, described in section 4.1, help to meet the need for more user-friendly and comprehensive access to OAI-compliant collections and resources. Second, through work led primarily by the DLF and NSDL in the US along with JISC in the UK, there are renewed efforts to create quality shareable metadata by promoting best practices, organizing training workshops, and "marketing" the value of metadata (refer to section 3.1). Issues of granularity are aided by recommendations to use enriched MODS metadata that describes objects more fully. Meanwhile the quantity of object-level data has mushroomed thus offering users with more coherent content. Third, concerns about providing users with a meaningful "context" if not fully realized, are increasingly remedied by improvements in aligning collections with object-level data and through new visualization and clustering techniques. Moreover there is a better understanding of both the potential and limitations of metadata-driven technical infrastructures. New digital architectures, such as implemented by the NSDL, emphasize relationships among resources (hence give "context") in which metadata plays an important but not singular or preeminent role.

5.3.2 Users and Uses

 "Users and uses" are frequently the starting point—rather than a by-product—of building distributed libraries. Studies such as "Use and Users of Digital Resources: A Focus on Undergraduate Education in the Humanities and Social Sciences" (Harley et al. 2006), (which is now being adapted to study the sciences by Alan Wolf and Flora McMartin), and JISC's "Disciplinary Differences" (Sparks 2005) offer a more refined articulation of faculty preferences and environmental constraints. Increasingly, user or persona scenarios are developed for a wide variety of purposes such as explaining the need for new technologies (Frumkin 2006b), evaluating repository platforms (Choudhury 2006), or creating new services (American West).[211] Further, virtually all of the services under review in this report have conducted at least one user study. The DLF Aquifer Services Institutional Survey Report (2006) found that most user evaluations by its members come at the point of introducing or updating a service, therefore, DLF

---

[211] See respective project sites:

Digital Library Service Registries, Use Studies:

http://wiki.library.oregonstate.edu/confluence/display/DLSRW/RegistryUseCases/

A Technology Analysis of Repositories and Services:

https://wiki.library.jhu.edu/display/RepoAnalysis/ProjectRepository/

American West, see persona scenarios:

http://www.cdlib.org/inside/projects/amwest/

Aquifer hopes to develop a model for the "persistent assessment" of how digital resources are used and integrated into various service environments.

5.3.3 Managing Digital Rights and Digital Content Preservation

A second broad direction identified in the 2003 report, "finding solutions to digital rights management and digital content preservation" is now being addressed on multiple fronts through numerous high-profile initiatives, a few of which are highlighted here predominantly in relationship to the services under review and OAI-PMH.[212] Inspired in part by the RoMEO Project (Rights MEtadata for Open archiving, described in the 2003 DLF report), the Open Archives Initiative released specifications in May 2005 documenting how to express rights at the record-level and at the repository and set aggregation levels, "Conveying rights expressions about metadata in the OAI-PMH framework."[213] It guides both data and service providers in the optimal way to create and harvest rights management metadata. Directories of journal and publisher's policies regarding self-archiving—another outgrowth of the RoMEO Project—help librarians and authors to determine publishing and distribution options (described in section 4.1). In spring 2006, to provide immediate access to embargoed journal articles, EPrints.org announced the release of a "Request eprint" button in its software to enable interested readers to request authors to supply them with an email full-text version of a restricted access article. In response, DSpace made a similar add-on available, called "RequestCopy" (http://wiki.dspace.org/RequestCopy).

In the vast realm of digital preservation, the PREMIS (PREservation Metadata: Implementation Strategies) Working Group, a team of 30 experts from five countries jointly sponsored by OCLC and RLG, completed its work and released its products, including the Data Dictionary for Preservation Metadata issued in June 2005. The dictionary and associated XML schema are now maintained under the auspices of the Library of Congress (LC) (http://www.loc.gov/standards/premis/). The LC Digital Preservation Web site provides up-to-date news about NDIIPP (National Digital Information Infrastructure and Preservation Program, http://www.digitalpreservation.gov/). From here, readers can obtain the latest information about Technical Infrastructure developments, Collaborative Collection Development Partnerships, Research Awards, E-depot for e-journals (Portico), States Initiatives, and Organization Alliances. In the UK, the Digital Curation Centre, established in 2004, is the focal point for research, training, and publication about digital preservation. Finally, the tutorial designed by Cornell University Library, "Digital Preservation Management: Implementing Short-term Strategies for Long-term Problems," won the 2004 publication award from the Society of American Archivists. It

---

[212] For background about Digital Rights Management (DRM) in its much broader context, refer to http://en.wikipedia.org/wiki/Digital_rights_management/

[213] Available from http://www.openarchives.org/OAI/2.0/guidelines-rights.htm.

provides an excellent introduction to the issues along with a listing of other resources and publications (in need of update as of mid-2006).[214]

5.3.4 Building Personal Libraries and Collaborative Workspaces

Among the services under review, there are certainly efforts towards this goal exemplified by NSDL's incorporation of more interactive and social networking features or the Sheet Music Consortium's provision to create personal collections. The Scholars Box (described in section 4.5) is designed to facilitate "interoperability across four intersecting domains of interoperability: educational technology, library services, desktop tools, and social software." Collex, under development at the University of Virginia, is an "open-source collections- and exhibits-builder designed to aid humanities scholars working in digital collections or within federated research environments like NINES" (described in section 4.4.11). Perseus Digital Library hopes to implement "a distributed editing environment whereby users may correct error, comment on topics, create custom commentaries, user guides, discuss issues with other users, and personalize the Perseus experience."

Community-building is an objective of many of the sites and they encourage collaboration through such activities as peer review, implementation of editorial boards, or integrating user comments about resources. With noticeable advancements towards integration of resources into personal work spaces, many of these services are poised to deliver new user-driven functionality in the not-distant future. However, none of them have yet to attain the level of what ARTstor (http://www.artstor.org/) has to offer in terms of providing users with the tools to manage and integrate externally-created and hosted digital images with personal and institutional collections (Marmor 2006).

"Save yourself! Free resources for organising, maintaining and sharing the fruits of your web searches," (Bates 2006) reviews personalization and social-networking tools offered by generic and niche services.[215]

5.3.5 Putting Digital Libraries in the Classroom and Digital Objects in the Curriculum

A major finding of the Center for Studies in Higher Education (UC, Berkeley) about faculty's use of digital resources in undergraduate education, bears reiterating: "…they simply do *not* mesh with faculty members pedagogies" (Harley et al. 2006, 49). If ARTstor represents a superior model of community-responsiveness in developing tools, content, and services that do coincide with instructional practice, there are, nevertheless,

---

[214] The tutorial is available from http://www.library.cornell.edu/iris/tutorial/dpm/eng_index.html.
[215] Available from http://www.freepint.com/issues/160306.htm.

examples among many of the services considered in this report where similar efforts are in the planning, if not implementation, stage.

The NSF/NSDL-funded, Instructional Architect (http://ia.usu.edu/), for example, "allows you to find, use, and share learning resources from the National Science Digital Library (NSDL) and the Web in order to create engaging and interactive educational web pages." Services such as MERLOT have signed agreements with several e-learning platforms, although what they have to share is metadata about learning resources, not the objects themselves. NEEDS architecture supports cataloging in full IEEE-LOM compliant metadata and it reports working on a more extensive cataloging interface to leverage that ability and to provide users with richer metadata of the learning objects. It is modifying its "authority lists" or vocabulary to conform to agreed upon standards, such as currently being done for "learning resource type." Built through a collaborative design process, DLESE Teaching Boxes are "classroom-ready instructional units created by collaboration between teachers, scientists, and designers. Each box helps to bridge the gap between educational resources and how to implement them in the classroom. The Teaching Boxes contain materials that model scientific inquiry, allowing teachers to build classroom experiences around data collection and analysis from multiple lines of evidence, and engaging students in the process of science." (http://www.teachingboxes.org/).

Repository technology platforms are also seeking solutions to achieve interoperability with e-learning systems. Browsing and searching for content in DSpace via the open source Sakai learning environment (http://sakaiproject.org/) is already possible and DSpace is now examining integration with Moodle (http://moodle.org/) and Blackboard (http://wiki.dspace.org/SakaiIntegration/).

Finally, several influential studies and a collaborative JISC/NSF project are worth noting.

- "Interoperability between Library Information Services and Learning Environments – Bridging the Gaps," a Joint White Paper written by Neil McLean and Clifford Lynch on behalf of the IMS Global Learning Consortium and the Coalition of Networked Information (May 10, 2004) scopes out library interactions with the e-learning space, examines issues related to different conceptualizations of repositories and stewardship, and provides an overview of the IMS Digital Repositories Interoperability Framework. http://www.imsglobal.org/digitalrepositories/CNIandIMS_2004.pdf
- "Digital Library Content and Course Management Systems: Issues of Interoperation," The Report of a Study Group co-chaired by Dale Flecker and Neil McLean under the aegis of the Digital Library Federation (July 2004), designed a model of instructional "workflow" practices, applied the model to use cases, analyzed what services and practices repository owners should consider when designing their offerings, and created an extensive "checklist" of

service requirements and best practices for repositories.
http://www.diglib.org/pubs/cmsdl0407/cmsdl0407.htm#summary

- JORUM, *a free online repository service for teaching and support staff in UK Further and Higher Education Institutions, helping to build a community for the sharing, reuse and repurposing of learning and teaching materials*, has produced a series of useful reports surveying: international e-learning repository initiatives and commercial systems; technical frameworks; open source learning object repository systems; digital rights management; and digital preservation issues.
http://www.jorum.ac.uk/

- The Digital Libraries in the Classroom Programme (scheduled to end in July 2006) is an international program jointly funded by JISC and the National Science Foundation (NSF), developed *to bring about significant improvements in the learning and teaching process in certain disciplines within higher education in the US and UK, through bringing emerging technologies and readily available digital content into mainstream educational use.* Its four funded projects include:
  - o The Spoken Word – led by Glasgow Caledonian University and Michigan in partnership with the BBC exploring the use of digital audio in the humanities, http://www.spokenword.ac.uk/;
  - o DialogPlus – a partnership between the University of Southampton, the University of Leeds, Penn State and the University of California, Santa Barbara, working in the Geography discipline, http://www.dialogplus.org/;
  - o DIDET – a partnership between the University of Strathclyde and Stanford University working in the design engineering discipline, http://www.didet.ac.uk/;
  - o DART – a partnership between the London School of Economics and Columbia University in the discipline of Anthropology, http://www.columbia.edu/dlc/dart/.

5.3.6 Promoting Excellence

Mechanisms to promote excellence are often built into the structure of the services under review, for example by establishing submission routines to ensure author credibility (e.g. arXiv's user endorsement system), guidelines for metadata compliance (e.g., OLAC's metadata report card evaluation system), creating peer-review systems (e.g., BEN), setting up editorial boards (e.g., NINES), and distributing awards for excellence (e.g., NEEDS Premier Award for courseware). Projects like the "Cream of Science" in the Netherlands meet the twin goal of fulfilling institutional repositories while showcasing the work of top scholars (http://www.creamofscience.org/).

The Certificate of the DINI German Initiative for Network Information (described in section 4.1) serves as quality filter for institutional data providers by supporting minimum standards and recommendations. The Certificate is awarded to the repository

after review by a distributed group of experts. As of May 2006 there are nineteen DINI-certified document servers (http://www.dini.de/dini/zertifikat/zertifiziert.php).

At a systemic level, the standards and best practices discussed in this report (e.g., those promoted by the DLF, NSDL, NISO, etc.) are intended to improve the overall quality and interoperability of distributed libraries. Geared towards preservation and long-term sustainability of digital resources, the RLG-NARA "Audit Checklist for the Certifying Digital Repositories" (draft of August 2005), builds on the Open Archival Information System (OAIS) Reference Model (ISO 14721) adopted in 2002 and related high-level articulation of the attributes and responsibilities for trusted, reliable sustainable digital repositories (RLG and OCLC 2002). Efforts to move this proposal forward to implementation are underway at the Center for Research Libraries, through a grant funded by The Andrew W. Mellon Foundation, Participating archives include the Royal Library of the Netherlands, Portico, the Inter-university Consortium for Social Science Research (ICPSR), and LOCKSS (Lots of Copies Keep Stuff Safe).[216]

5.4 The Pulse in 2006

5.4.1 Acceptance of OAI-PMH and Growth in Adoption

This report leaves little doubt that the Open Archives Initiative Protocol for Metadata Harvesting has witnessed remarkable international adoption and growth since 2003. In case after case, the aggregations under review recorded sizeable gains in the number of records available via their services, frequently noting this growth as one of their three most significant accomplishments. More than 1,000 OAI-compliant archives are active across at least 46 countries with an estimated seven million links to full digital object representation. OAI modules have become a standard feature in institutional repository software and e-publishing platforms—whether open source or commercial. This trend is perhaps best exemplified by the highly acclaimed HighWire Press, which has a well-established tradition of offering free access to a large proportion of its journal article database, but débuted as a registered OAI data provider in 2006, starting with Oxford University Press journals (http://openarchive.highwire.org/).[217] Adoption is likely to accelerate as more countries view OAI implementation as a fast-track to bring increased visibility to indigenous scholarship.

---

[216] Information about the CRL's Certification of Digital Archives program is available from http://www.crl.edu/content.asp?l1=13&l2=58&l3=142.

[217] Although it makes no mention of OAI per se, Péter Jacsó's April 2006 review of the Oxford Journal Collection ("Péter's Digital Reference Shelf" GaleNet) provides extensive analysis of the extent of its OA content as well as its accessibility via HighWire Press. Available from http://www2.hawaii.edu/~jacso/.

Along with the good news come two cautionary tales. First, the bulk of OAI items with full-object representation come from a limited number of countries and sites. Data derived from both ROAR and OAIster suggest that half of all records are supplied by repositories in the United States, United Kingdom, and Germany; and that the largest top 20 services constitute 70 percent or more of all records (see Appendix 04). The influence of a handful of repositories is undeniable, for example, CiteSeer, PubMed Central, arXiv, and American Memory. In contrast to these services, the average deployment has fewer than 12,000 items and the median hovers around 500 records. Overall, thematic- or discipline-based archives have been more effective thus far in attracting content than university-based repositories. Aside from IRs built around research agencies such as the U.S. Office of Scientific and Technical Information (OSTI) or CERN in Switzerland, university IRs appear to have relatively few full-text resources. The largest IR in OAIster, Demetrius Australia National University Institutional Repository, had 42,000 items as of March 2006. As discussed in this report, the situation will change dramatically if and when more "self-archiving" mandates are invoked by institutions, funding agencies, or through national legislation.

Secondly, there is growing awareness of the limitations of OAI-PMH and the Dublin Core metadata standard that "underpin much of the current repository activity" along with a call to develop a model and mechanisms to handle "complex objects held in repositories … in a more fully automated and interoperable way" (Heery and Powell 2006, 18). A meeting sponsored and supported by Microsoft, The Andrew W. Mellon Foundation, the Coalition for Networked Information, the Digital Library Federation, and JISC explored these issues with the intention of reaching "agreement on the nature and characteristics of a limited set of core, protocol-based repository interfaces (REST-full and/or SOAP-based Web services) that allow downstream applications to interact with heterogeneous repositories in an efficient and consistent manner; compile a concrete list of action items aimed at fully specifying, validating and implementing such repository interfaces; and devise a timeline for the specification, validation and implementation of such repository interfaces" (OAI News, http://www.openarchives.org/news/news2.html#InterOp).

5.4.2 Interoperability in an International Framework

Strategic planning for interoperability takes place increasingly in the international arena. Open access converges with open source platforms and open standards that are worked out with international input. Whether the discussion revolves around e-learning, e-research or Web publishing, many projects, principles, platforms, and policies bridge national borders. Examples abound throughout this report from bi-national partnerships such as the JISC (UK) and DEST (Australia) e-Framework, to transnational movements like the Berlin Declaration on Open Access. DSpace, EPrints.org, and Fedora are international communities of practice. NISO's Metasearch Initiative has involved more than 60 individuals from five countries (Hodgson, Pace and Walker 2006). Systems to

measure use and research impact are attempting to synchronize efforts across national borders.

Service providers surveyed in this report identify both accomplishments and challenges of an international dimension. OLAC is actively engaged in establishing best practices in digital language documentation, calling out for better language identification in metadata. The Library of Congress is taking a leadership role with OCLC and the Deutsche Bibliothek to harmonize millions of people's names across catalogs (name authority control) through the creation of the Virtual International Authority File. At the same time, LC seeks more tools to support multilingual search and display, as it moves to create the Global Gateway. MERLOT has joined up with GLOBE (Global Learning Objects Brokered Exchange) international consortium and offers federated searches across the European, Ariadne, and Australian, EdNA learning object collectives. CiteSeer and arXiv have established mirror sites on an international basis. The CERN Document Server translates its services in 14 languages and the NDLTD Union catalog represents ETD content in more than 25 languages.

In comparison to many other countries where higher education strategic planning and funding of networked infrastructures and digital services are coordinated by centralized agencies, the situation in the United States is much more decentralized involving a variety of public and private funding agencies (e.g., NSF, IMLS, Mellon, Hewlett), higher education coalitions and federations (e.g., CNI, Educause) and library-related entities (e.g., DLF, OCLC/RLG, the Library of Congress, ARL). The California Digital Library, Digital Library Federation, National Science Digital Library, and National Digital Information Infrastructure and Preservation Program stand out as four different models used in the US to pursue high-level, multi-dimensional digital agendas across a wide sector of stakeholders. Still, in contrast to concerted investigations in the UK and the Netherlands, for example, in the realm of connecting digital repositories to national and pan-European networks, the United States higher education community lacks a widely accepted organizational vehicle for developing parallel frameworks.

5.4.3 Sustainability and Funding—Ubiquitous Concerns

The most common challenge and resource requirement cited by survey respondents revolved around funding and staffing. This concern cut across all services irrespective of status, business model, organizational structure, community of practice, or subject domain. Survey respondents noted:

- arXiv: "staff time and money"
- OAIster: "need to recruit a programmer"
- OLAC: "sponsorship for maintaining core services; guidance on long-term funding sources other than research agencies"

- NSDL: "lack of funding to offer more teacher workshops in how to use NSDL…" and "lack of a well-funded corporate and foundation outreach program to diversify sustainability options"
- NEEDS: "sustainability planning"
- MERLOT: "high demand but limited resources"
- Cornucopia: "funding"
- Heritage West: "sustained funding"
- Aquifer: "outside funding that would allow dedicated project staff; support for service model development to evaluate organizational effectiveness and to plan for sustainability"
- SouthComb: "sustainability of service: managing the transition from project to ongoing program"
- Perseus: "meeting needs of growing audience with limited resources, including providing adequate user support; ability to maintain current services while also implementing research agendas"
- NINES: "funding to sustain the developing infrastructure; funding to move paper-based journals that want to become part of the NINES project to online operations"
- INFOMINE: "continued funding of programmers and metadata specialists"

This collective « cri de cœur » begs for widespread, cross-sector dialogue, training, and strategic action, drawing on the findings of Zorich (2003), Bishoff and Allen (2004), Berkman (2004) and the work of NSDL's Sustainability Standing Committee, especially its Sustainability Matrix and Vignettes (http://sustain.comm.nsdl.org/).[218] Addressing funding and sustainability dovetails with the need to improve the marketing of these services and integrate them into existing scholarly information systems.

5.4.4 Next Generation Service Characteristics

Culled from survey responses, these "next generation" service features—many of which are under development if not already deployed—are grouped according to what they offer users and advantages they will bring to service providers.[219]

From the user perspective, next generation services:

- Will be developed with a better understanding of user needs and reflect the "scholars' voice."

---

[218] As this report goes to press, a study commissioned by JISC has been released that investigates various business models to support "Linking UK Repositories" (Swan and Awre, 2006).

[219] For a recent articulation of what "The Open Research Web" will offer, refer to Shadbolt et al. in Jacobs, forthcoming 2006.

- Will be trusted and preferred, offering valuable services beyond generic search engines.
- Will be fully integrated with other scholarly resources and embedded in scholarly environments, thus more widely used in learning, teaching, and research.
- Will offer mechanisms for scholars to disseminate research findings and to navigate the literature by citation linking and impact rankings. Will enable scholars to measure usage and impact. Similarly, will permit instructors to assess the pedagogic value of digital resources used in different teaching settings.
- Will enable easy discovery of appropriate metasearch portals and the ability to dynamically select resources to be metasearched at the moment of query.
- Will support sound and video recording; natural language queries in multi-lingual environments; data and text mining; dynamic clustering; interpretation and analysis; side-by-side comparison, manipulation, and re-use of digital objects in local environments.
- Will offer more push technologies, community-building tools (threaded discussion forums, blogs, newsletters), collaborative tools (share baskets, alerts, annotations, comments, reviews), and interactive features.
- Will leverage multiple online and face-to-face interactions as repeat users become contributors in a timely and transparent way.

From the service provider perspective, next generation services:

- Will have more easy to use tools that allow collections to become OAI compliant along with mechanisms for service providers to better assess OAI conformance and communicate shortcomings efficiently with data providers.
- Will have more automated, robust, flexible open source tools for metadata creation, normalization and enrichment. Barriers to participation will be lower, while quality becomes higher.
- Will have the means to automatically ingest digital objects, along with bulk-loading tools (OAI-based at first) to ingest applicable, already-cataloged collections quickly. Tools to mine data from existing catalogs and authority files will improve ingested records.
- Will reduce administrative time and labor through better facilities and easier submission processes.
- Will offer collaborative tools that allow trusted others to contribute to or edit site records remotely; and improved object relational management systems;
- Will have automated means to cleanse metadata and manage duplicate records.
- Will speed up indexing as there is more support for OAI-PMH flow control.
- Will have standards and mechanisms in place to measure and share usage and impact values across repositories.
- Will lead to improved search and retrieval systems through automated, dynamic classifiers and semantic clustering techniques. Tools will be in place to support

surfacing topical cohesiveness across highly heterogeneous aggregated collections.

- Will deploy user quality metrics for metasearch systems, making it possible to customize search and retrieval to specific user needs. Improved classifiers and crawlers will help to scale with the increase of scholarly resources.

- Will enjoy widespread adoption of search protocols (e.g., SRU, MXG) by vendors and aggregators. Digital library service registries will enable effective machine-to-machine and direct end-user access to requested resources.

- Will enable deep sharing through experimentation with aggregation other than metadata harvesting, resulting in the capacity to move digital objects from domain to domain, along with the ability to modify and re-deposit them in a different location in the process.

- Will feature new cluster and file systems that help to automate building and deploying digital libraries, making it easy for users to install and utilize them.

References

Agogino, A.M. 2004. Enhancing Interoperability of Collections and Services. Final Report, December 2004. NSF Award DUE-0127580. Available from http://best.me.berkeley.edu/%7Eaagogino/papers/Final_Report_SMETE.pdf

Almasy, Edward. 2005. Tools for Creating Your Own Resource Portal: CWIS and the Scout Portal Toolkit. *Library Trends* 53(4): 620-636.

American Council on Learned Societies. 2005. *Draft Report of the ACLS Commission on Cyberinfrastructure for Humanities and Social Sciences* (November 24, 2005). Available from http://www.acls.org/cyberinfrastructure/acls-ci-public.pdf.

Arms, Carol R. 2003. Available and Useful: OAI at the Library of Congress. *Library Hi Tech*, *21*(2): 129-139 [DOI:10.1108/07378830310491899] Available from http://memory.loc.gov/ammem/techdocs/libht2003.html#exploit

Bailey, Charles W., Jr. 2005. *Open Access Bibliography: Liberating Scholarly Literature with E-Prints and Open Access Journals*. Washington DC: Association of Research Libraries. Available from http://www.digital-scholarship.com/.

Bailey, Charles W., Jr. 2006. What Is Open Access? In Jacobs, Neil, ed. *Open Access: Key Strategic, Technical and Economic Aspects*. Oxford: Chandos Publishing. Preprint available from http://www.digital-scholarship.com/cwb/WhatIsOA.pdf

Bailey-Hainer, Brenda and Richard Urban. 2004. The Colorado Digitization Program: A Collaboration Success Story. *Library Hi Tech* 22(3): 254-262.

Bates, Mary Ellen. 2006. Tips Article: "Save Yourself! Free Resources for Organising, Maintaining and Sharing the Fruits of your Web Searches." *Freepint Newsletter* 202 (March 16). Available from http://www.freepint.com/issues/160306.htm.

Berkman, Paul Arthur. 2004. Sustaining the National Science Digital Library. *Project Kaleidoscope Newsletter*, Vol. IV: What Works, What Matters, What Lasts. Available from http://www.pkal.org/documents/Vol4SustainingTheNSDL.cfm

Bishoff, Liz and Nancy Allen. 2004. *Business Planning for Cultural Heritage Institutions.* Washington DC: Council on Library and Information Resources. Available from http://www.clir.org/pubs/reports/pub124/contents.html/

Blinco, Kerry et al. 2004. Trends and Issues in E-Learning Infrastructure Development. A White Paper for alt-i-lab 2004. Prepared on behalf of DEST (Australia) and JISC-CETIS (UK). Version 2, July 19, 2004. Available from http://www.jisc.ac.uk/uploaded_documents/Altilab04-infrastructureV2.pdf

Blinco, Kerry and McLean. 2004. A 'Cosmic' View of Repositories Space: Wheel of Fortune. University of Southern Queensland Available from http://www.rubric.edu.au/extrafiles/wheel/

Blinco, Kerry. 2006. The JISC-DEST Framework: Integrating Everything. PowerPoint presentation at Open Repositories, Sydney, Australia (February 2, 2006). Available from http://www.apsr.edu.au/Open_Repositories_2006/conference_program.htm#thursday

Bloomberg, Jason. 2006. SOA Governance and the Butterfly Effect. Zapthink.com. (January 24). Available from http://www.zapthink.com/report.html?id=ZAPFLASH-2006124.

Brody, T., Harnad, S. and Carr, L. 2005. Earlier Web Usage Statistics as Predictors of Later Citation Impact. *Journal of the American Association for Information Science and Technology (JASIST)* Eprint available from http://eprints.ecs.soton.ac.uk/10713/.

Brogan, Martha L. 2003. *A Survey of Digital Library Aggregation Services.* Washington DC: Digital Library Federation. Available from http://www.diglib.org/pubs/brogan/

California Digital Library. 2004. *National Science Digital Library: Focus Groups and Market Assessment, Final Report.* Prepared by Alex Wright with others. (1 July 2004). Available from www.cdlib.org/inside/projects/metasearch/nsdl/nsdl_assessmentfindings.pdf

Campbell, Debbie. 2005. The ARROW Discovery Service 'Public Funding, Public Knowledge, Public Access.' National Library of Australia *Gateways* 46 (August). Available from http://www.nla.gov.au/pub/gateways/archive/76/Campbell-ArrowProject.html

Campbell, Debbie. 2004. How the Use of Standards Is Transforming Australian Digital Libraries. *Ariadne 41* (October). Available from http://www.ariadne.ac.uk/issue41/campbell/intro.html

Cervone, Frank H. 2004. The Repository Adventure. *Library Journal*. June 1 2004. Available from http://www.libraryjournal.com/article/CA421033.html

Chan, Leslie. 2004. Supporting and Enhancing Scholarship in the Digital Age: The Role of Open-Access Institutional Repositories. Available from http://eprints.rclis.org/archive/00002590/01/Chan_CJC_IR.pdf (not copied?)

Chan, Leslie, Barbara Kirsop and Subbiah Arunachalam. 2005. Open Access Archiving: The Fast Track to Building Research Capacity in Developing Countries. First published as part of the SciDev.Net Web Site. (November). Available from http://www.scidev.net/ms/openaccess/

Chang, Amy et al. 2004. I Came, I Found It, I Used It, and It Made a Difference. BiosciEdNet (BEN). [n.p.]: American Society of Microbiology, 8 pp. Available from http://www.biosciednet.org/project_site/BEN_Survey_Article_October_2004.pdf

Choudhury, Sayeed. 2006. A Technology Analysis of Repositories and Services. CNI Spring 2006 Task Force Meeting hand-out. Alexandria, Virginia, April 3-4 2006. Available from http://www.cni.org/tfms/2006a.spring/abstracts/Handouts/CNI_Results_Choudhury.doc. Project wiki accessible from https://wiki.library.jhu.edu/display/RepoAnalysis/ProjectRepository/.

Christenson, Heather and Roy Tennant. 2005. Integrating Information Resources: Principles, Technologies, and Approaches. In partial fulfillment of National Science Foundation Award #0333710 (August 15). Available from http://www.cdlib.org/inside/projects/metasearch/nsdl/nsdl_report2.pdf

Chudnov, Dan and Jeremy Frumkin. 2005. Digital Library Dialtone: Bootstrapping with Auto-discovery and Service Registries. PowerPoint presented at the DLF Fall Forum, Charlottesville, Virginia, November 7, 2005. Available from http://www.diglib.org/forums/fall2005/presentations/ockhambof1105_files/frame.htm

Cole, Timothy W. and Sarah L. Shreeves. 2004. Search and Discovery across Collections: The IMLS Digital Collections and Content Project. *Library Hi Tech 22*(3): 307-322.

Cole, Timothy W. 2005a. OAI-PMH Repositories: Quality Issues Regarding Metadata and Protocol Compliance: Part II Shareable Metadata. Presented at CERN workshop on *Innovations in Scholarly Communication (OAI4)*, Geneva, Switzerland, 20 October 2005. PowerPoint slides available from http://imlsdcc.grainger.uiuc.edu/OAI4_Tutorial_1_URLs.htm#shareable

Cole, Timothy W. 2005b. Rationale for Interoperable Metadata. DLF Fall Forum, Charlottesville, Virginia, November 7-9, 2005. Available from http://www.diglib.org/forums/fall2005/presentations/cole1105.htm

Cole, Timothy W. 2006. OAI-PMH in Practice: Lessons Learned from the IMLS Digital Collections & Content Project. Middle East Digital Library Workshop, Bibliotheca Alexandrina, 15-17 January 2006, Available from http://www.sis.pitt.edu/~egyptdlw/presentations.html

Cole, Timothy W. and Thomas G. Habing. 2006. The University of Illinois OAI-PMH Data Provider Registry. Presentation at NSF/DLF/JISC/UKOLN Digital Library Service Registry Workshop. National Science Foundation, Arlington, Virginia, March 23, 2006. Available from http://gita.grainger.uiuc.edu/registry/Cole-OAIRegistry.ppt

Coleman, Anita and Joseph Roback. 2005. Open Access Federation for Library and Information Science: dLIST and DL-Harvest. *D-Lib Magazine*. *11*(12) (December 2005). Available from http://www.dlib.org/dlib/december05/coleman/12coleman.html

Coutts, Brian E. and Cheryl LaGuardia. 2006. Best Reference 2005. *Library Journal* (April 15). Available from http://www.libraryjournal.com/article/CA6321696.html.

Crane, Gregory. 2006. Featured Collection: The Perseus Digital Library. *D-Lib Magazine,* 12,3 (March). Available from http://www.dlib.org/dlib/march06/03featured-collection.html. doi:10.1045/march2006-featured.collection.

Crow, Raym. 2002. The Case for Institutional Repositories: A SPARC Position Paper. Association of Research Libraries, Scholarly Publishing & Academic Resources Coalition. Available from http://www.arl.org/sparc/IR/ir.html.

Custard, Myra and Tamara Sumner. 2005. Using Machine Learning to Support Quality Judgments. *D-Lib Magazine,* 11, 10 (October). Available from http://www.dlib.org/dlib/october05/custard/10custard.html. doi:10.1045/october2005-custard

Davison, Stephen, Cynthia Requardt and Kristine Brancolini. 2003. A Specialized Open Archives Initiative Harvester for Sheet Music: A Project Report and Examination of Issues. Presented at ISMIR 2003, October 26-30. Available from http://ismir2003.ismir.net/papers/Davison.PDF

Dempsey, Lorcan. *Lorcan Dempsey's Weblog On Libraries, Services and Networks*. Available from http://orweblog.oclc.org/.

Dempsey, Lorcan. 2004. Libraries, Digital Libraries and Digital Library Research. PowerPoint presentation at the European Conference on Digital Libraries, University of Bath, September 12-17, 2004. Available from http://www.ecdl2004.org/presentations/dempsey/l-dempsey.ppt.

Dempsey, Lorcan. 2005. The User Face that Isn't. May 15, 2005. Available from http://orweblog.oclc.org/archives/000667.html.

Dempsey, Lorcan et al. 2004-05. Metadata Switch: Thinking About Some Metadata Management and Knowledge Organization Issues in the Changing Research and Learning Landscape. Forthcoming in *LITA Guide to E-Scholarship* (working title), ed. Debra Shapiro. Available from http://www.oclc.org/research/publications/archive/2004/dempsey-mslitaguide.pdf (PDF:824K/25pp.)

Dewatripont, Mathias et al. 2006. *Study on the Economic and Technical Evolution of the Scientific Publications Market in Europe.* Final Report, commissioned by the Directorate-General of Research, European Commission (January). Available from http://ec.europa.eu/research/science-society/pdf/scientific-publication-study_en.pdf.

Diekema, Amy and Holly Devaul. 2006. Computer Assisted Standard and Alignment. Audio PowerPoint presentation at NSDL Tool Time, March 16, 2006. Archived video playback available from http://commserv.comm.nsdl.org/tooltime/2006-03-16/lib/playback.html.

Digital Library Federation. 2005. The Distributed Library: OAI for Digital Library Aggregation. OAI Scholars Advisory Panel Meeting, Washington DC, June 20-21, 2005. Available from http://www.diglib.org/architectures/oai/imls2004/OAISAP05.htm.

Digital Library Federation. DLF Aquifer Services Working Group. 2006. DLF-Aquifer Services Institutional Survey Report. (March 9). Available from http://www.diglib.org/aquifer/SWGisrfinal.pdf

Digital Library Federation and the National Science Digital Library. OAI and Shareable Metadata Best Practices Working Group. 2005. OAI Best Practices. Available from http://oai-best.comm.nsdl.org/

- *Best Practices for OAI Data Provider Implementations*. Available from http://oai-best.comm.nsdl.org/cgi-bin/wiki.pl?DataProviderPractices
- *Best Practices for Shareable Metadata* http://oai-best.comm.nsdl.org/cgi-bin/wiki.pl?PublicTOC

Dine, Brooke. 2004. PubMed Central. PowerPoint presentation given at the Medical Library Association Conference, May 2004. Available from: http://www.nlm.nih.gov/pubs/techbull/ja04/theater_ppt/pmc.ppt.

Dobratz, Susanne and Astrid Schoger. 2005. Digital Repository Certification: A Report from Germany. *RLG DigiNews* 15, 9 (October). Available from http://www.rlg.org/en/page.php?Page_ID=20793#article3

Dominguez, Magaly Báscones. 2005. Applying Usage Statistics to the CERN E-journals Collection: a Step Forward, *High Energy Physics Libraries Webzine*, 11 (August). Available from http://library.cern.ch/HEPLW/11/papers/4/ .

Dunsire, Gordon. 2005. Harvesting Institutional Resources in Scotland Testbed (HaIRST). Final Report. Available from http://hairst.cdlr.strath.ac.uk/documents/HaIRST-FAIR-FP.pdf

Fabos, Bettina., ed. 2005. The Commercialized Web: Challenges for Libraries and Democracy. *Library Trends 53*(4): 519-698.

Foster, Nancy Fried and Susan Gibbons. 2005. Understanding Faculty to Improve Content Recruitment for Institutional Repositories. *D-Lib Magazine 11*(1) (January). Available from http://www.dlib.org/dlib/january05/foster/01foster.html.

Foulonneau, Muriel. 2005. CIC Metadata Portal: Project Status. PowerPoint presented at the Big Ten Center, Chicago, December 12, 2005.

Foulonneau, Muriel and Timothy W. Cole. 2004. CIC-OAI Project Recommendations for Dublin Core Metadata Providers," Version 1.0 (06/18/2004). University of Illinois. Available from http://cicharvest.grainger.uiuc.edu/dcguidelines.asp.

Foulonneau, Muriel, Thomas G. Habing, and Timothy W. Cole. 2006. Automated Capture of Thumbnails and Thumbshots for Use by Metadata Aggregation Services. *D-Lib Magazine 12*(1), (January). Available from http://www.dlib.org/dlib/january06/foulonneau/01foulonneau.html

Foulonneau, Muriel et al. 2005. Using Collection Descriptions to Enhance an Aggregation of Harvested Item-Level Metadata. PowerPoint presented at JCDL, June 2005. Available from http://cicharvest.grainger.uiuc.edu/presentations/jcdl.ppt

Foulonneau, Muriel et al. 2006. The CIC Metadata Portal: A Collaborative Effort in the Area of Digital Libraries. Forthcoming in Sci Tech Lib. 2006. Preprint available from http://cicharvest.grainger.uiuc.edu/documents/PreprintCIC-OAI-ScienceTechnologyLibraries.pdf

Frumkin, Jeremy. 2006a. The Need for A Digital Library Service Registry. *OCLC Systems & Services; International Digital Library Perspective 22*(1): 23-25.

Frumkin, Jeremy. 2006b. Registry Use Cases. Developed for the Digital Library Service Registry Workshop, co-sponsored by DLF and JISC, March 22-23, 2006, Washington DC. Available from http://wiki.library.oregonstate.edu/confluence/display/DLSRW/RegistryUseCases/.

Gargiulo, Paola, Susanna Mornati, Ugo Contino, and Zeno Tajoli. 2005a. PLEIADI, A Portal Solution for Scholarly Literature. In Dobreva, Milena and Engelen, Jan, Eds. Proceedings ELPUB2005. *From Author to Reader: Challenges for the Digital Content Chain: Proceedings of the 9th ICCC International Conference on Electronic Publishing*. Leuven-Heverlee (Belgium). (May 2005): 277-281. Available from http://eprints.rclis.org/archive/00004442/

Gargiulo, Paola, Susanna Mornati, Ugo Contino, and Zeno Tajoli. 2005b. A User-Centred Portal for Search and Retrieval of Open-Access Italian Scholarly Literature: The PLEIADI Project. In *Proceedings Open Culture : Accessing and Sharing Knowledge*, Milano (Italy). (June 2005). Available from http://eprints.rclis.org/archive/00004403/01/Pleiadi_AICA2005_rev.pdf

Gentil-Beccot, Anne. 2006. 2005, the Year CERN Ran for Open Access. *High Energy Physics Libraries Webzine*, *12* (February). Available from http://library.cern.ch/HEPLW/12/papers/3/

Ghosh, S. B. 2006. Open Access and Institutional Repositories – A Developing Country Perspective: A Case Study of India. Preprint for World Library and Information Congress: 72nd IFLA General Conference and Council 20-24 August 2006, Seoul, Korea. Available from http://www.ifla.org/IV/ifla72/papers/157-Ghosh_Das-en.pdf.

Giersch, Sarah et al. 2004. If You Build It, Will They Come ? Participant Involvement in Digital Libraries. *D-Lib Magazine 10*, 7/8 (July/August). Available from http://www.dlib.org/dlib/july04/giersch/07giersch.html.

Goldenberg-Hart, Diane. 2004. Libraries and Changing Research Practices: A Report of the ARL/CNI Forum on E-Research and Cyberinfrastructure. *ARL Bimonthly Report*, 237 (December). Available from http://www.arl.org/newsltr/237/cyberinfra.html.

Gonçalves, Marcos André et al. 2004. Streams, Structures, Spaces, Scenarios, Societies (5S): A Formal Model for Digital Libraries. *ACM Transactions on Information Systems 22*(2) (April): 270-312.

Giustini, Dean and Eugene Barsky. 2005. A Look at Google Scholar, PubMed, and Scirus: Comparisons and recommendations. *JCHLA/JABSC 26*: 85-89. Available from: http://www.slais.ubc.ca/courses/libr538f/04-05-wt2/giustini_barsky.pdf

Gray, Cindy et al. 2005. The Online ETD Tutorial: Supporting Students' Transition from Traditional to ETD. Presentation at the NDLTD annual conference, Australia 2005. Available from http://adt.caul.edu.au/etd2005/papers/065Gray.pdf.

Guenther, Rebecca. 2005. Rich Descriptive Metadata in XML: MODS as a Metadata Scheme. Presentation at "MODS, MARC and Metadata Interoperability," at ALA Annual 2005. Available from http://www.loc.gov/standards/mods/presentations/ala2005-mods_files/frame.htm

Habing, Tom. 2005. UIUC's Role: Registry of OAI Data Providers. PowerPoint presented at the DLF Fall Forum, Charlottesville, Virginia, November 8, 2005. Available from http://www.diglib.org/forums/fall2005/presentations/habing-oaigrant1105_files/frame.htm

Habing, Thomas G., Timothy W. Cole, and William H. Mischo. 2004. Developing a Technical Registry of OAI Data Providers. Presented at the ECDL 2005 conference, University of Bath, England, September 15, 2004. Available as a Post-print from Lecture Notes in Computer Science, Springer-Verlag from http://gita.grainger.uiuc.edu/registry/thabing_ecdl2004.doc

Hagedorn, Kat. 2005a. IMLS Grant: University of Michigan's Role. PowerPoint presented at the DLF Fall Forum, Charlottesville, Virginia, November 8, 2005. Available from http://www.diglib.org/forums/fall2005/presentations/hagedorn-oaigrant1105_files/frame.htm

Hagedorn, Kat. 2005b. And Now for Something Completely Different. . . Informal Collaboration. University of Michigan School of Information Internet Public Library (IPL) Class, 28 November 2005, Ann Arbor, MI. Available from http://oaister.umdl.umich.edu/o/oaister/IPL05_Hagedorn.ppt

Halbert, Martin. 2003. Findings from the Metascholar Projects: AmericanSouth and MetaArchive. *Proceedings of the Workshop on Applications of Metadata Harvesting in Scholarly Portals*. Atlanta: MetaScholar Initiative at Emory University, October 24, 2003. Available from http://www.metascholar.org/pdfs/MetaScholarFindingsProceedings.pdf

Halbert, Martin. 2005. DLF IMLS OAI Best Practices Project: Training Component. PowerPoint presented at the DLF Fall Forum, Charlottesville, Virginia, November 8, 2005. Available from http://www.diglib.org/forums/fall2005/presentations/halbert-oaigrant1105_files/frame.htm

Hanson, Katherine and Bethany Carlson. 2005. Effective Access: Teachers' Use of Digital Resources in STEM Teaching. Gender, Diversities, and Technology Institute, Education Development Center, Inc. Available from http://www2.edc.org/gdi/publications_SR/EffectiveAccessReport.pdf.

Hardy, Rachel et al. 2005. Open Access Citation Information. Final Report – Extended Version. JISC Scholarly Communications Group. (September), 105 p. Download PDF from JISC Scholarly Communications Group site, available from http://www.jisc.ac.uk/index.cfm?name=jcie_scg.

Harley, Diane et al. 2006. Use and Users of Digital Resources: A Focus on Undergraduate Education in the Humanities and Social Sciences. Center for Studies of Higher Education, University of California, Berkeley (April). Available from http://cshe.berkeley.edu/research/digitalresourcestudy/

Harnad, Stevan et al. 2003. Mandated Online RAE CVs Linked to University Eprint Archives: Enhancing UK Research Impact and Assessment. *Ariadne 35* (April). Available from http://www.ariadne.ac.uk/issue35/harnad/

Harnad, S. & Brody, T. 2004. Comparing the Impact of Open Access (OA) vs. Non-OA Articles in the Same Journals. *D-Lib Magazine 10*(6) (June). Available from http://www.dlib.org/dlib/june04/harnad/06harnad.html

Harnad, S., Brody, T., Vallieres, F., Carr, L., Hitchcock, S., Gingras, Y, Oppenheim, C., Stamerjohanns, H., & Hilf, E. 2004. The Access/Impact Problem and the Green and Gold Roads to Open Access. http://dx.doi.org/10.1016/j.serrev.2004.09.013. *Serials Review 30*(4). http://eprints.ecs.soton.ac.uk/10209/01/impact.html#bib15

Healey, Michael. 2000. An Analytical Model of Collections and Their Catalogues. A study carried out on behalf of UKOLN. Oxford. Available from http://www.ukoln.ac.uk/metadata/rslp/model/amcc-v31.pdf.

Healey, Michael 2005. Users and Information Resources: An Extension of the Analytical Model of Collections and Their Catalogues into Usage and Transactions. A study carried out on behalf of UKOLN. Oxford, 2nd rev. issue, July-November 2005. Available from http://www.ukoln.ac.uk/cd-focus/model-ext/CD2-principles-v2-2.pdf.

Heery, Rachel and Sheila Anderson. 2005. Digital Repositories Review. UKOLN, AHDS. (February 19). Available from http://www.jisc.ac.uk/uploaded_documents/digital-repositories-review-2005.pdf.

Heery, Rachel and Andy Powell. 2006. Digital Repositories Roadmap: Looking Forward. UKOLN, Eduserv Foundation. (April). Available from http://www.jisc.ac.uk/uploaded_documents/rep-roadmap-v15.doc

Hey, Tony et al. 2006. Augmenting Interoperability Across Scholarly Repositories. JCDL Panel Abstract, June 11-15, 2006, Chapel Hill, North Carolina. Available from http://msc.mellon.org/Meetings/Interop/follow-up/jcdlpanel_abstract.pdf

Hiom, Debra. 2006. Retrospective on the RDN. *Ariadne 47* (April). Available from http://www.ariadne.ac.uk/issue47/hiom/intro.html.

Hitchcock, Stephen. [September 15, 2004 – present]. The Effect of Open Access and Downloads ('Hits') on Citation Impact: A Bibliography of Studies. OpCit Project. Available from http://opcit.eprints.org/oacitation-biblio.html.

Hodgson, Cynthia, Andrew Pace, and Jenny Walker. 2006. NISO Metasearch Initiative Targets Next Generation of Standards and Best Practices. *Against the Grain 18*(1) (February): 79-82. Available from http://www.niso.org/pdfs/ReprintNISOv18-1.pdf.

Hubbard, William. 2005. "OpenDOAR The Directory of Open Access Repositories." Presented at OAI4. Available from http://www.sherpa.ac.uk/documents/OAI4OpenDOAR_pub.ppt

Hughes, Baden. 2004. Metadata Quality Evaluation: Experience from the Open Language Archives Community. *Lecture Notes in Computer Science*. 3334: 320-329. Available from http://eprints.unimelb.edu.au/archive/00001408/

Hughes, Baden and Amol Kamat, 2005. A Metadata Search Engine for Digital Language Archives**.** *DLib Magazine* 11(2), (February). Available from http://www.dlib.org/dlib/february05/hughes/02hughes.html.

Hunter, Philip. 2005. OAI and OAI-PMH for Absolute Beginners: a Non-Technical Introduction. PowerPoint presentation at the CERN workshop on Innovations in Scholarly Communication (OAI4) (Geneva, October 20-22, 2005). Self-archived January 30, 2006. Available from http://eprints.rclis.org/archive/00005512/.

Hutchings, Paul and Ilana Levin. 2006. NIH Author Postings: A Study to Assess Understandings of, and Compliance with, NIH Public Access Policy. Conducted on behalf of the Publishing Research Consortium (February). Available from http://www.alpsp.org/news/NIH_authorpostings_report.pdf

Jacobs, Neil. 2006a. International Workshop on E-Theses. *Ariadne 46* (February). Available from http://www.ariadne.ac.uk/issue46/e-theses-rpt/

Jacobs, Neil, ed. 2006b. *Open Access: Key Strategic, Technical and Economic Aspects.* Oxford, England: Chandos Publishing.

Jacsó, Péter. 1993. A Proposal for Database "Nutrition and Ingredient" Labeling. Guest Editorial, "The Linear File," *Database* (February): 7-9.

Jacsó, Péter. 2004. CiteBaseSearch, Institute of Physics Archive, and Google's Scholarly Archive. "Péter's Picks and Pans."*Online* 28, 5 (Sept/Oct): 57-60. Available from http://www2.hawaii.edu/~jacso/.

Jacsó, Péter. 2005a. CiteSeer. Péter's Digital Reference Shelf. *GaleNet* (November). Available from http://www2.hawaii.edu/~jacso/.

Jacsó, Péter. 2005b. Visualizing Overlap and Rank Differences Among Web-wide Search Engines: Some Free Tools and Services. *Online Information Review* 29 (5): 554-60.

Jascó, Péter. 2006. The Oxford Journals Collection. Péter's Digital Reference Shelf. *Galenet* (April). Available from http://www2.hawaii.edu/~jacso/.

Jin, Hai. [2004]. ChinaGrid Overview. Available from http://unpan1.un.org/intradoc/groups/public/documents/apcity/unpan016934.pdf

Jones, Paul. 2005. Strategies and Technologies of Sharing in Contributor-Run Archives. *Library Trends* 53(4): 651-662.

Jones, Richard. 2004. The Tapir: Adding E-Theses Functionality to DSpace. *Ariadne* 41 (October). Available from http://www.ariadne.ac.uk/issue41/jones/.

Jones, Richard, et al. 2006. *The Institutional Repository*. Oxford, England: Chandos Publishing.

Kastens, K. 2005. The DLESE Community Review System: Gathering, Aggregating, and Disseminating User Feedback about the Effectiveness of Web-based Educational Resources. *Journal of Geoscience Education* 53: 37-43. Available from http://www.nagt.org/files/nagt/jge/abstracts/Kastens_v53n1.pdf

Kastens, Kim et al. 2005. Questions & Challenges Arising in Building the Collection of a Digital Library for Education: Lessons from Five Years of DLESE. *D-Lib* 11(11). Available from http://www.dlib.org/dlib/november05/kastens/11kastens.html

Kastens, Kim A. and Neil Holzman. 2006. The Digital Library for Earth System Education Provides Individualized Reports for Teachers on the Effectiveness of Educational Resources in Their Own Classrooms. *D-Lib* 12, 1 (January 2006). Available from http://www.dlib.org/dlib/january06/kastens/01kastens.html.

Khan, Haseebulla M., Kurt Maly and Mohammad Zubair. 2005. Similarity and Duplicate Detection System for an OAI Compliant Federated Digital Library. Proceedings of the 9th European Conference, ECDL 2005, Vienna, Austria, September 18-23, 2005. in *Lecture Notes in Computer Science*, "Research and Advanced Technology for Digital Libraries," vol. 2652: 531-532.

Kelly, Brian, Amanda Closier, and Debra Hiom. 2005. Gateway Standardization: A Quality Assurance Framework for Metadata. *Library Trends* 53(4): 637-650.

Kennan, Mary Anne et al. 2005. ADT ProQuest Collaboration : A Case Study of a Library/Vendor Alliance. Available from http://adt.caul.edu.au/etd2005/papers/095Kennan.pdf.

Koelling, Jill and Mark Shelstad. 2006. Collaborative Digitization Projects. PowerPoint presentation at the Midwest Archives Conference, April 27-29, 2006, Normal, Illinois. Available from http://www.cdpheritage.org/cdp/presentations/documents/CollaborativeDigitizationProgramsMAC06.pdf.

Kott, Katherine et al. 2005. DLF Aquifer: Bringing Collections to Light. PowerPoint presented at the DLF Fall Forum, Charlottesville, Virginia on November 8, 2005. Available from http://www.diglib.org/forums/fall2005/presentations/aquifer1105_files/frame.htm

Kott, Katherine et al. 2006. DLF Aquifer: Phase 1 Accomplishments. PowerPoint presented at the DLF Spring Forum Austin, Texas on April 11, 2006. Available from http://www.diglib.org/forums/spring2006/presentations/aquifer0406_files/frame.htm

Kraan, Wilbert and Jon Mason. 2005. Issues in Federating Repositories: Report on the First International CORDRA™ Workshop. *D-Lib Magazine* 11,3 (March), Available from http://www.dlib.org/dlib/march05/kraan/03kraan.html.

Lagoze, Carl et al. 2002. "The Open Archives Initiative Protocol for Metadata harvesting, Version 2.0." (June). Available from http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm

Lagoze, Carl et al. 2005. What Is a Digital Library Anymore, Anyway? Beyond Search and Access in the NSDL. *D-Lib Magazine* 11,11 (November). Available from http://www.dlib.org/dlib/november05/lagoze/11lagoze.html#n3

Lagoze, Carl et al. 2006a. Metadata Aggregation and "Automated Digital Libraries": A Retrospective on the NSDL Experience. Submission to JCDL. Preprint available from http://www.arxiv.org/ftp/cs/papers/0601/0601125.pdf.

Lagoze, Carl et al. 2006b. Representing Contextualized Information in the NSDL. Preprint available from http://www.arxiv.org/ftp/cs/papers/0603/0603024.pdf .

LaGuardia, Cheryl. 2006. ResearchNow. *Library Journal* (February 1). Available from http://www.libraryjournal.com/article/CA6299837.html.

Landis, Bill. 2005. Collaborative Metadata Aggregations: CDL's American West Experience + A Whole Lot of Unanswered Questions. PowerPoint presented at the DLF Fall Forum, Charlottesville, Virginia, November 8, 2005. Available from http://www.diglib.org/forums/fall2005/presentations/landis1105_files/frame.htm.

Landis, Bill. 2006. Go Fish! Experiments with Topical Metadata Enhancement in the American West Project. DLF Spring Forum, Austin, Texas (April 11). Available from http://www.diglib.org/forums/spring2006/presentations/landis0406_files/frame.htm.

Lavoie Brian, Lorcan Dempsey, and Lynn Silipigni Connaway. 2006. Making Data Work Harder. *Library Journal.* (January 15).Available from http://www.libraryjournal.com/article/CA6298444.html

Liu, Xiaoming et al. 2005. Lessons Learned with Arc, an OAI-PHM Service Provider. *Library Trends*. 53(4): 590-603.

Lossau, Norbert. 2006. Existing Infrastructures to Bring Cultural Heritage Online, and New Needs For Such Infrastructures. Presentation at "Open Access to Cultural Heritage," Golm/Potsdam, Germany, 29 March 2006. Available from http://berlin4.aei.mpg.de/presentations/Lossau_OA06.pdf.

Lutz, Marilyn. 2005. Featured Collection: The Maine Music Box. *D-Lib Magazine.* 11, 3 (March). Available from http://www.dlib.org/dlib/march05/03featured-collection.html.

Lynch, Clifford A. 2003. Institutional Repositories: Essential Infrastructure for Scholarship in the Digital Age. *ARL Bimonthly Report*, 226 (February): 1-7. Available from http://www.arl.org/newsltr/226/ir.html.

Lynch, Clifford A. and Joan K. Lippincott. 2005. Institutional Repository Deployments in the United States as of Early 2005. *D-Lib Magazine* 11, 9 (September). Available from http://www.dlib.org/dlib/september05/lynch/09lynch.html.

Lyon, Liz and Simon Coles. 2004. eBank UK: Linking Research Data, Scholarly Communication and Learning. Presentation at All Hands Meeting: eScience Broadening the Horizon, Nottingham, England sponsored by the National eScience Centre, September 2, 2004. Available from http://www.nesc.ac.uk/events/ahm2004/presentations/237.ppt.

MacLeod, Roddy and Malcolm Moffat. 2005. Engineering Digital Repositories Landscape Analysis, and Implications for PerX. Version 1.0 (10/11/05). Available from http://www.icbl.hw.ac.uk/perx/analysis.htm.

Marlino, M. R., T. R. Sumner, and M. J. Wright, 2004. Geoscience Education and Cyberinfrastructure. Report of a workshop sponsored by the National Science Foundation (NSF), April 19-20. Boulder: Digital Library for Earth System Education (DLESE) Program Center; University Corporation for Atmospheric Research (UCAR), 43p. Available from http://www.dlese.org/documents/reports/GeoEd-CI.html

Marmor, Max. 2006. Six Lessons Learned: An (Early) ARTstor Retrospective. *RLG DigiNews* 10, 2 (April). Available from http://www.rlg.org/en/page.php?Page_ID=20916#article0. McCown, Frank, Johan Bollen and Michael L. Nelson. 2005. Evaluation of the NSDL and Google for Obtaining Pedagogical Resources. *Lecture Notes in Computer Science* 3652, 344-355.

McCown, Frank, Xiaoming Liu, Michael L. Nelson, Mohammad Zubair. 2005. Search Engine Coverage of the OAI-PMH Corpus. *IEEE Internet Computing* LA-UR-04-9158. Available from http://library.lanl.gov/cgi-bin/getfile?LA-UR-05-9158.pdf.

McGann, Jerome. 2005. Like Living on the Nile. IVANHOE, A User's Manual. *Literature Compass* 2, VI 149 (2005): 1-27.

McGann, Jerome and Bethany Nowviskie. 2005. NINES: A Federated Model for Integrating Digital Scholarship. White Paper (September):1-40. Available from http://www.nines.org/about/9swhitepaper.pdf.

McKiernan, G. 2005. E-profile: Scirus: For Scientific Information Only. *Library Hi Tech News*, 22,3 (Mar):18-25.

McLean, Neil. 2004. The Ecology of Repository Services: A Cosmic View. PowerPoint presentation at ECDL (European Conference on Research and Advanced Technology for Digital Libraries), University of Bath, September 12-17, 2004. Available from http://www.ecdl2004.org/presentations/mclean/.

McLean, Neil and Clifford Lynch. 2004. Interoperability between Library Information Services and Learning Environments – Bridging the Gaps. A Joint White Paper on behalf of the IMS Global Learning Consortium and the Coalition of Networked Information. (May 10). Available from http://www.imsglobal.org/digitalrepositories/CNIandIMS_2004.pdf.

Mimno, David, Gregory Crane and Alison Jones. 2005. Hierarchical Catalog Records: Implementing a FRBR Catalog. *D-Lib Magazine*, 11,10 (October). Available from http://www.dlib.org/dlib/october05/crane/10crane.html. doi:10.1045/october2005-crane.

Mitchell, Steven. 2005. Collaboration Enabling Internet Resource Collection-Building Software and Technologies. *Library Trends* 53(4): 604-619.

Mitchell, Steven. 2006. Automated Metadata Generation and New Resource Discovery Software and Services. PowerPoint Presentation at IMLS Webwise 2006. Available from http://datafountains.ucr.edu/IMLS2006presentation.ppt

Mitchell, Steven, Julie Mason and Lori Pender. 2004. Enabling Technologies and Service Designs for Collaborative Internet Collection Building. *Library Hi Tech* 22, 3: 295-306.

Moffat, M. 2006. 'Marketing' with Metadata - How Metadata Can Increase Exposure and Visibility of Online Content. PerX project, Version 1.0 8th March 2006. Available from http://www.icbl.hw.ac.uk/perx/advocacy/exposingmetadata.htm#nine.

NASA. Scientific & Technical Information Program Office. 2005. Persistent Unique Identifiers Being Added to NASA STI. *STI Bulletin Online.* (July). Available from http://www.sti.nasa.gov/Pubs/Bulletin/05julypub/05julypub.pdf

NASA. Scientific & Technical Information Program Office. 2006. Richer Search Experience Coming for NASA STI Public Collection. *STI Bulletin Online.* (January). Available from http://www.sti.nasa.gov/Pubs/Bulletin/06janpub/06janpub.pdf

National Information Standards Organization. 2004. A Framework of Guidance for Building Good Digital Collections. 2nd edition. Available from http://www.niso.org/framework/Framework2.html.

National Institutes of Health. Department of Health and Human Services. 2006. *Report on the NIH Public Access Policy*. Submitted by Elias A. Zerhouni, Director, NIH, (January).

National Science Foundation. NSF Cyberinfrastructure Council. 2006. *Cyberinfrastructure Vision for 21st Century Discovery.* Version 7.1. July 20, 2006. Available from http://www.nsf.gov/od/oci/ci-v7.pdf

Nelson, Michael L. and Johan Bollen. 2005. "If You Harvest arXiv.org, Will They Come?" Fifth ACM/IEEE Joint Conference on Digital Libraries (JCDL 2005), June 7 - 11 (Denver, CO), p. 393. 1-58113-876-8/05/006.

Nelson, Michael L., Johan Bollen, JoAnne Rocker and Calvin Mackey. 2004. User Evaluation of the NASA Technical Report Server Recommendation Service, Sixth ACM CIKM International Workshop on Web Information and Data Management (WIDM 2004), Washington D.C., USA, November 12-14, 2004. ACM 2004. Available from http://ntrs.nasa.gov/NTRSrecommendation1004.pdf

Nelson, Michael L., JoAnne Rocker and Terry L. Harrison. OAI and NASA's Scientific and Technical Information. *Library Hi Tech*, 21 (2):140-150.

Noruzi, Alireza. 2005. Google Scholar: The New Generation of Citation Indexes. *Libri* 55:170-180.

Nowviskie, Bethany. 2005. COLLEX: Semantic Collections and Exhibits for the Remixable Web. (November). Available from http://www.nines.org/about/Nowviskie-Collex.pdf.

OCLC Online Computer Library Center, Inc. 2005. *Perceptions of Libraries and Information Resources.* A Report to the OCLC Membership. Contributions by Cathy De Rosa et al. Dublin, Ohio: OCLC. Available from http://www.oclc.org/reports/2005perceptions.htm**.**

OCLC/RLG PREMIS Working Group. 2004. Implementing Preservation Repositories for Digital Materials: Current Practice and Emerging Trends in the Cultural Heritage Community. Report by the joint OCLC/RLG Working Group Preservation Metadata: Implementation Strategies (PREMIS). Dublin, Ohio: OCLC Online Computer Library Center, Inc. Available from http://www.oclc.org/research/projects/pmwg/surveyreport.pdf

Open Archives Initiative. 2002. Specification for an OAI Static Repository and an OAI Static Repository Gateway. Protocol Version 2.0 of 2002-06-14. Document Version 2004/04/23T15:17:00Z http://www.openarchives.org/OAI/2.0/guidelines-static-repository.htm

Paynter, G., 2005. Developing Practical Automatic Metadata Assignment and Evaluation Tools for Internet Resources. JCDL, Denver. Available from http://ivia.ucr.edu/projects/publications/Paynter-2005-JCDL-Metadata-Assignment.pdf.

Peek, Robin. 2005. RCUK Free for All. *Information Today*. 22, 8 (September): 17-18.

Pepe, Alberto, Jean-Yves Le Meur and Tibor Šimko. 2006. Dissemination of scientific results in High Energy Physics: the CERN Document Server vision. Presented at Computing in High Energy and Nuclear Physics (Mumbai, February 13-17, 2006). Available from http://indico.cern.ch/getFile.py/access?contribId=216&sessionId=5&resId=0&materialId=paper&confId=048. Conference Web site http://www.tifr.res.in/%7Echep06/

Petricek, Vaclav et al. 2005. A Comparison of On-Line Computer Science Citation Databases. ECDL 2005, *Lecture Notes in Computer Science* 3653: 438-449. Available from http://citeseer.ist.psu.edu/petricek05comparison.html

Phipps, Jon, Diane Hillmann and Gordon Paynter. 2004. Orchestrating Metadata Enhancement Services: Introducing Lenny. http://www.arxiv.org/ftp/cs/papers/0501/0501083.pdf.

Plutchak, T. Scott. 2005. The Impact of Open Access. *JMLA: Journal of the Medical Library Association. 93*(4) October 2005: 419-421. Available from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1250314

Poe, Felicia with contributions from Jane Lee. 2005. American West Broad Topic Category Naming: Survey Findings & Recommendations. Prepared for the American West Project Team, (May). Available from http://www.cdlib.org/inside/assess/evaluation_activities/docs/2005/survey_May2005_report.pdf.

Powell, Andy. 2005. A 'Services-Oriented' View of the JISC Information Environment. UKOLN, University of Bath. (November). Available from http://www.ukoln.ac.uk/distributed-systems/jisc-ie/arch/soa/jisc-ie-soa.pdf.

Proudman, Vanessa. 2006. Nereus: One International Subject-based Repository Meeting Two Needs - Libraries and Researchers Collaborate. Delivered at 8th International Bielefeld

Conference, Bielefeld (Germany). Available from http://conference.ub.uni-bielefeld.de/2006/docs/presentations/proudman_biconf06_final.ppt.

Renda, M. Elena and Umberto Straccia. 2005. A Personalized Collaborative Digital Library Environment: a Model and an Application. *Information Processing & Management*. 41,1 (January): 5-21.

RLG and NARA. 2005. An Audit Checklist for the Certification of Trusted Digital Repositories. Draft for Public Comment. Mountain View, CA: RLG, August 2005. Available from http://www.rlg.org/en/pdfs/rlgnara-repositorieschecklist.pdf.

RLG and OCLC. 2002. Trusted Digital Repositories: Attributes and Responsibilities. Mountain View, CA: RLG, May 2002. Available from http://www.rlg.org/en/pdfs/repositories.pdf.

Roosendaal, Hans and Peter Geurts. 1997. Forces and Functions in Scientific Communication: an Analysis of their Interplay. Cooperative Research Information Systems in Physics, August 31—September 4 1997, Oldenburg, Germany. Available from http://www.physik.uni-oldenburg.de/conferences/crisp97/roosendaal.html.

Sadeh, Tamar. 2006. Google Scholar Versus Metasearch Systems. *HEP Libraries Webzine* 12 (March). Available from http://library.cern.ch/HEPLW/12/papers/1/#ref-02.

Sale, Arthur. 2006. The Impact of Mandatory Policies on ETD Acquisition. *D-Lib Magazine* 12, 4 (April). Available from http://www.dlib.org/dlib/april06/sale/04sale.html.

Seaman, David. **2005a.** OAI For Digital Library Aggregation: An IMLS Grant-funded DLF Project. PowerPoint presented at the DLF Fall Forum, Charlottesville, Virginia, November 8, 2005. Available from
http://www.diglib.org/forums/fall2005/presentations/cole1105.htm

Seaman, David. 2005b. Podcast Interview at CNI 2005 Fall Task Force Meeting. Phoenix, Arizona http://connect.educause.edu/featured_content/mpasiewicz/an_interview_with_dlfs_executive_director_david_seaman/1712

Shadbolt, Nigel et al. 2006. The Open Research Web. In Jacobs, Neil, ed. *Open Access: Key Strategic, Technical and Economic Aspects*. Oxford: Chandos Publishing, 2006. Preprint available from http://www.arxiv.org/ftp/cs/papers/0601/0601125.pdf.

Shreeves, Sarah L. et al. 2005. Current Developments and Future Trends for the OAI Protocol for Metadata Harvesting. *Library Trends* 53(4): 576-589.

Smith, MacKenzie. 2006. Transition to a Broader Participation: Experience from the DSpace Project. PowerPoint Presentation at Open Repositories conference, Sydney, Australia (February 3, 2006). Available from
http://www.apsr.edu.au/Open_Repositories_2006/conference_program.htm#thursday

Sparks, Sue. 2005. JISC Disciplinary Differences Report. London: Rightscom Ltd. Available from http://www.jisc.ac.uk/index.cfm?name=schol_comms_reports.

Sreekumar, M.G. 2006. Open Access Landscaping in India: Building Institutional Repositories (IRs) Using 'DSpace.' PowerPoint Presentation at Open Repositories conference, Sydney, Australia (February 1, 2006). Available from http://www.apsr.edu.au/Open_Repositories_2006/conference_program.htm#wednesday.

Straccia, Umberto and Costantino Thanos. 2004. An Open Collaborative Virtual Archive Environment. *International Journal on Digital Libraries* 4 (1): 23-24, DOI: 10.1007/s00799-003-0063-.

Suber, Peter. *Open Access News*. Weblog available from http://www.earlham.edu/~peters/fos/fosblog.html.

Suber, Peter. *The SPARC Open Access Newsletter*. Available from http://www.earlham.edu/~peters/fos/.

Sumner, Tamara et al. 2004. A Web Service Interface for Creating Concept Browsing Interfaces. *D-Lib Magazine* 10, 11 (November). Available from http://www.dlib.org/dlib/november04/sumner/11sumner.html.

Swan, Alma and Chris Awre. 2006. Linking UK Repositories: Technical and Organisational Models to Support User-Oriented Services Across Institutional and Other Digital Repositories. Scoping Study Report. (June). Available from http://www.jisc.ac.uk/uploaded_documents/Linking_UK_repositories_report.pdf

Sequeira E. 2005. PubMed® Links to Author Manuscripts in PubMed Central®. *NLM Techical Bulletin*, 345 (July-August): e3.

Swan, Alma et al. 2005. Developing a Model for E-prints and Open Access Journal Content in UK Further and Higher Education. *Learned Publishing* 18, 1: 25-40. Available from http://eprints.ecs.soton.ac.uk/11000/

Swan, Alma and Sheridan Brown. 2004a. Authors and Open Access Publishing. *Learned Publishing* 17(3) 219-224. http://cogprints.org/4123/

Swan, Alma and Sheridan Brown. 2004b. JISC/OSI Journal Authors Survey Report. http://cogprints.org/4125/

Swan, Alma and Sheridan Brown. 2005. Open Access Self-Archiving: An Author Study. Key Perspectives Ltd. Technical Report, Joint Information Systems Committee (JISC), UK FE and HE Funding Councils. (May). Available from http://cogprints.org/4385/ or http://www.keyperspectives.com

Tansley, Robert. 2006. The China Digital Museum Project. PowerPoint presentation at Open Repositories conference, Sydney, Australia (January 31, 2006). Available from http://www.apsr.edu.au/Open_Repositories_2006/conference_program.htm#tuesday

Teng, Xia et al. 2005. Best Practices in the Design, Development and Use of Courseware in Engineering Education. Presented at the 25th ASEE/IEEE Frontiers in Education Conference, October 19-22, 2005 in Indianapolis, Indiana, Session T1A-1. Available from http://www.needs.org/smete/public/about_smete/publications/FIE05/FIE_bestpractice_15.pdf

Tennant, Roy. [n.d.] Specifications for Metadata Processing Tools. California Digital Library. Available from http://www.cdlib.org/inside/projects/harvesting/metadata_tools.htm.

Tennant, Roy. [2004a] Bitter Harvest: Problems & Suggested Solutions for OAI-PMH Data & Service Providers. California Digital Library. Available from http://www.cdlib.org/inside/projects/harvesting/bitter_harvest.html.

Tennant, Roy. 2004b. A Bibliographic Metadata Infrastructure for the 21st Century. *Library Hi Tech* 22, 2: 175-181, Available from http://roytennant.com/metadata.pdf .

Turner, Chris. 2004. Cornucopia: An Open Collection Description Service. *Ariadne* 40 (July). Available from http://www.ariadne.ac.uk/issue40/turner/

Turner, Chris. 2005. Cornucopia. UKOLN Collection Description Focus. Case Study 4, January 2005. Available from http://www.ukoln.ac.uk/cd-focus/case-studies/cdf-casestudy4-cornucopia.pdf

United Kingdom. HM Treasury. 2006. Science and Innovation Investment Framework 2004-2014: Next Steps. (March). Available from http://www.hm-treasury.gov.uk/media/1E1/5E/bud06_science_332.pdf.

University Library Regensburg. 2006. Electronic Journals Library: Annual Report 2005. Regensburg: Germany, April 2006. Available from http://rzblx1.uni-regensburg.de/ezeit/anwender/Jahresbericht_EZB_2005_engl.pdf.

University of California Libraries. Bibliographic Services Task Force. 2005. Rethinking How We Provide Bibliographic Services for the University of California. Final Report: December 2005. Available from http://libraries.universityofcalifornia.edu/sopag/BSTF/Final.pdf.

Van de Sompel, Herbert et al. 2004. Rethinking Scholarly Communication: Building the System that Scholars Deserve. *D-Lib Magazine* (September). Available from http://www.dlib.org/dlib/september04/vandesompel/09vandesompel.html

Van de Sompel, Herbert et al. 2005. aDORe: A Modular, Standards-Based Digital Object Repository *The Computer Journal*, 48 (5): 514 - 535. doi:10.1093/comjnl/bxh114.

van Westrienen, Gerard. 2005. Completed Questionnaires: Country Update on Academic Institutional Repositories. Making the Strategic Case for Institutional Repositories, CNI-JISC-SURF Conference; Amsterdam, 10-11 May 2005. (August). Available from http://www.surf.nl/download/country-update2005.pdf.

van Westrienen, Gerard and Clifford A. Lynch. 2005. Academic Institutional Repositories: Deployment Status in 13 Nations as of Mid 2005. *D-Lib Magazine* 11, 9 (September). Available from http://www.dlib.org/dlib/september05/westrienen/09westrienen.html.

van Veen, Theo and Bill Oldroyd. 2004. Search and Retrieval in The European Library: A New Approach. *D-Lib Magazine* 10, 2 (February). Available from http://www.dlib.org/dlib/february04/vanveen/02vanveen.html

Vise, David A. 2005. World Digital Library Planned: Library of Congress Envisions Collection to Bridge Cultures. *The Washington Post*. (November 22): A27. Available from http://www.washingtonpost.com/wp-dyn/content/article/2005/11/21/AR2005112101428.html.

Warner, Simeon. 2005. The arXiv: Fourteen Years of Open Access Scientific Communication. In M. Halbert, ed. *Free Culture and the Digital Library Symposium Proceedings.* Atlanta: MetaScholar Initiative at Emory University, October 14, 2005, 56-68.

Warner, Simeon and Michael Nelson. 2003. Report on the Metadata Harvesting Workshop at JCDL 2003. ACM SIGIR Forum, 37,2 (Fall).

Walters, William H. 2006. Institutional Journal Costs in an Open Access Environment. Helen A. Gaser Library, Millersville University, Millersville, PA. Available from http://www.library.millersville.edu/public_html/walters/journal_costs.pdf .

Waters, Donald. 2001. The Metadata Harvesting Initiative of the Mellon Foundation. *ARL Bimonthly Report* 217 (August). Available from http://www.arl.org/newsltr/217/waters.html

Weatherley, J. 2005. A Web Service Framework for Embedding Discovery Services in Distributed Library Interfaces. In *Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries* (Denver, CO, USA, June 07 - 11, 2005). JCDL '05. ACM Press, New York, NY, 42-43. DOI= http://doi.acm.org/10.1145/1065385.1065394

Whitney, Colleen and Peter Brantley. 2006. Project Briefing: The Melvyl Recommender Project. PowerPoint Presentation at CNI Spring 2006 Task Force Meeting. Arlington, Virginia, April 3-4, 2006. Available from http://www.cni.org/tfms/2006a.spring/abstracts/PB-whitney-melvyl.html.

Willinsky, John. 2006. *The Access Principle: The Case of Open Access to Research and Scholarship*. Cambridge, Massachusetts and London: The MIT Press.

Wright, Michael. 2004. GEON and DLESE: Building Educational CyberInfrastructure. PowerPoint Presentation from GEON NSF meeting. Available from http://www.geongrid.org/ahm/Wright_Final_2004.ppt.

Yeomans, Joanne. 2006. CERN's Open Access E-print Coverage in 2006 : Three Quarters Full and Counting. *High Energy Physics Libraries Webzine*, 12, (March). Available from http://library.cern.ch/HEPLW/12/papers/2/

Zia, Lee I. 2006. The NSF National Science, Technology, Engineering, and Mathematics Education Digital Library (NSDL) Program. *D-LIb Magazine,* 12, 3 (December). Available from http://www.dlib.org/dlib/march06/03inbrief.html.

Zorich, Diane M. 2003. A Survey of Digital Cultural Heritage Initiatives and Their Sustainability Concerns. Washington DC: Council on Library and Information Resources, pub 118 (June). Available from http://www.clir.org/pubs/abstract/pub118abst.html.

Appendix 1
Survey Respondents and Contacts

Alice M. AGOGINO
NEEDS
http://www.needs.org/
SMETE.org
http://www.smete.org/

Linda AKLI
BiosciEdNet (BEN) Collaborative
http://www.biosciednet.org/portal/

Jenny BENEVENTO
IMLS Digital Collections & Content
http://imlsdcc.grainger.uiuc.edu/

Paul BERKMAN
NSDL: National Science Digital Library
Sustainability Committee
http://sustain.comm.nsdl.org/

Steven BIRD
Open Language Archives
http://www.language-archives.org/

Tim BRODY
Citebase
http://citebase.eprints.org/
ROAR: Registry of Open Access
Repositories
http://archives.eprints.org/
Journal Policies—Self-archiving Policies
of Journals
http://romeo.eprints.org/

Vinod CHACHRA
VTLS: NDLTD Union Catalog Search
http://zippo.vtls.com/cgi-
bin/ndltd/chameleon/

Ann CHAPMAN
Cornucopia
http://www.cornucopia.org.uk/

Alex CHAUX
NSDL : National Science Digital Library
http://www.nsdl.org/

Heather CHRISTENSON
Metasearch Infrastructure Project
California Digital Library
http://www.cdlib.org/inside/projects/me
tasearch/

Anita COLEMAN
dL-Harvest: Digital Library of
Information Science & Technology **
http://dlharvest.sir.arizona.edu/

Stephen DAVISON
Sheet Music Consortium
digital.library.ucla.edu/sheetmusic/

Tom DEHN
NDLTD Catalog, OCLC
http://alcme.oclc.org/ndltd/servlet/OAI
Handler?verb=ListSets

Holly DEVAUL
DLESE: Digital Library for Earth System
Education
http://www.dlese.org/dds/index.jsp

Brooke DINE
PubMed Central
http://www.pubmedcentral.nih.gov/

Thomas DOWLING
OhioLink Worldwide ETD Search
http://search.ohiolink.edu/etd/world.cgi

Naomi DUSHAY
NSDL: National Science Digital Library
http://www.nsdl.org/

Curtis FORNADLEY
Sheet Music Consortium

digital.library.ucla.edu/sheetmusic/

Muriel FOULONNEAU
CIC Metadata Portal
http://cicharvest.grainger.uiuc.edu/

Edward FOX
NDLTD: Networked Digital Library of
Theses and Dissertations
http://www.ndltd.org/

Connie FRANKENFIELD
US States Implementing GILS**
Illinois State Library
http://states.gils.net/

Jeremy FRUMKIN
OCKHAM Initiative
http://www.ockham.org/

Paola GARGIULO
PLEIADI**
http://www.openarchives.it/pleiadi/

Lee GILES
CiteSeer
http://citeseer.ist.psu.edu/

Fred GULDEN
ARL Scholars Portal*
http://www.arl.org/access/scholarsporta
l/

Tom HABING
CIC Metadata Portal
http://cicharvest.grainger.uiuc.edu/
Grainger Engineering Library at
University of Illinois, Urbana-
Champaign
http://web.library.uiuc.edu/grainger/
IMLS Digital Collections and Content
imlsdcc.grainger.uiuc.edu/collections/
UIUC Digital Gateway to Cultural
Heritage Materials*
http://nergal.grainger.uiuc.edu/cgi/b/bib
/bib-idx
University of Illinois OAI-PMH Data
Provider Registry

http://gita.grainger.uiuc.edu/registry/

Kat HAGEDORN
OAIster
http://www.oaister.org/

Martin HALBERT
AmericanSouth*
http://www.americansouth.org/
MetaScholar
http://www.metascholar.org/
SouthComb
[not yet available]

Irma HOLTKAMP
Flashpoint, LANL*
http://library.lanl.gov/

Kaye HOWE
NSDL: National Science Digital Library
http://www.nsdl.org/

Bill HUBBARD
OpenDOAR
http://opendoar.org/
Publisher Copyright Policies & Self-
Archiving: SHERPA/RoMEO list
http://www.sherpa.ac.uk/romeo.php

Susan JESUROGA
NSDL: National Science Digital Library
http://www.nsdl.org/

Alison JONES
Perseus Digital Library
http://perseus.uchicago.edu/hopper/

Løtte JORGENSEN
DOAJ: Directory of Open Access
Journals
http://www.doaj.org/

Jill KOELLING
Heritage West
http://www.cdpheritage.org/collection/h
eritageWest.cfm

Katherine KOTT
DLF Aquifer
http://www.diglib.org/aquifer/

Bill LANDIS
The American West
http://www.cdlib.org/inside/projects/amwest/

Jean-Yves LE MEUR
CDS: CERN Document Server***
http://cds.cern.ch/

Susan LIEPA
AARLIN*
http://www.aarlin.edu.au/

Kimberly LIGHTLE
Formerly ENC Online*
http://goenc.com/

Calvin MACKEY
NTRS: NASA Technical Reports
Server***
http://ntrs.nasa.gov/

Gail MACMILLAN
NDLTD: Networked Digital Library for
Theses and Disserations
http://www.ndltd.org/

Kurt MALY
Arc
http://www.eng.odu.edu/arc/
ARCHON*
http://archon.cs.odu/edu/

Mary MARLINO
DLESE: Digital Library for Earth System
Education
http://www.dlese.org/dds/index.jsp

Julie MASON
INFOMINE
http://infomine.ucr.edu/

Jerome MCGANN
NINES
http://www.nines.org/

Elizabeth MILEWICZ
AmericanSouth*
http://www.americansouth.org/
MetaScholar
http://www.metascholar.org/
SouthComb
[not yet available]

Stephen MITCHELL
iVia
http://ivia.ucr.edu/
Data Fountains
http://ivia.ucr.edu/#DataFountains

Sharon MOMBRU
Scirus
http://www.scirus.com/

Carol MINTON MORRIS
National Science Digital Library (NSDL)
http://nsdl.org/

Michael NELSON
Formerly NTRS
http://ntrs.nasa.gov/

Michael NEUBERT
American Memory and other OAI
Collection
Library of Congress
http://memory.loc.gov/ammem/

John Mark OCKERBLOOM
The Online Books Page**
http://digital.library.upenn.edu/books/

John M. SAYLOR
NSDL: National Science Digital Library
http://www.nsdl.org/

David SEAMAN
DLF Digital Collections Registry
http://susanowo.grainger.uiuc.edu/DLF
COllectionsRegistry/browse/
DLF Portal
http://www.hti.umich.edu/i/imls/
DLF MODS Portal
http://www.hti.umich.edu/m/mods/

Edwin SEQUEIRA
PubMed Central
http://www.pubmedcentral.nih.gov/

Massimiliano SIMONCINI
DigitAlexandria**
http://www.bdaweb.net/

Katherine SKINNER
AmericanSouth*
http://www.americansouth.org/
MetaScholar
http://www.metascholar.org/
SouthComb
[not yet available]

Barbra SPERLING
MERLOT (Multimedia Educational
Resource for Learning and Online
Teaching)
http://www.merlot.org/Home.po

Umberto STRACCIA
Cyclades*
http://www.ercim.org/cyclades/

David SUNDVALL
NSDL: National Science Digital Library
http://www.nsdl.org/

Roy TENNANT
Metasearch Infrastructure Project
California Digital Library
http://www.cdlib.org/inside/projects/me
tasearch/

Christopher TURNER
Cornucopia
http://www.cornucopia.org.uk/

Simeon WARNER
arXiv.org
http://www.arxiv.org/

John WILLINSKY
Public Knowledge Project**
http://www.pkp.ubc.ca/

Caroline WILLIAMS
Intute
http://www.intute.ac.uk/development/

Jeff YOUNG
XTCat*
http://www.intute.ac.uk/development/

*Service excluded from study in 2006. Refer to Appendix 3.
**Service excluded or only partially covered.
***Did not provide feedback on draft text.

Appendix 2:
Other Specialists and Projects Consulted

Sayeed CHOUDHURY
Johns Hopkins University
http://ldp.library.jhu.edu/projects/repository/

Neil JACOBS
JISC Executive
http://www.jisc.ac.uk/

Péter JACSÓ
University of Hawaii
http://www2.hawaii.edu/~jacso/

Michael KAPLAN
PRIMO, Ex Libris
http://www.exlibrisgroup.com/

John HARRINGTON
AERADE
http://aerade.cranfield.ac.uk/

Marilyn LUTZ
Maine Music Box
http://mainemusicbox.library.umaine.edu/

Max MARMOR
ARTstor
http://www.artstor.org/

Paul A. S. NEEDHAM
AERADE
http://aerade.cranfield.ac.uk/

James PRINGLE
Thomson Scientific
http://scientific.thomson.com/

Jeff RIEDEL
Digital Commons/bepress
ProQuest
http://www.proquest.com/products_umi/digitalcommons/

Barbara ROCKENBACH
ARTstor
http://www.artstor.org/

Martin SCHEUPLEIN
EZB : Elektronische
Zeitschriftenbibliothek
http://rzblx1.uni-regensburg.de/ezeit/

David STERN
Yale University Science Libraries
http://www.library.yale.edu/science/services/

Donald J. WATERS
The Andrew W. Mellon Foundation
http://www.mellon.org/

Gary WIGGINS
Indiana University, School of Informatics
http://www.informatics.indiana.edu/

Appendix 3
2003 Services Excluded in 2006 Report

AARLIN: the Australian Academic & Research Library Network

http://www.aarlin.edu.au/

Goal: National Portal Framework

Status: Moved from project phase to an operating system and achieved its goals by December 2004. Continues as a priority with 12 member libraries deploying Ex Libris with Metalib and SFX modules.

- Summary of achievements reported on 12/9/2004.
  http://www.caul.edu.au/caul-doc/caul20042aarlin.pdf
- Strategic directions 2006-07 available from
  http://www.aarlin.edu.au/about.shtml

AmericanSouth.org

http://americansouth.org/

Goal: Scholar-designed portal prototype service

Status: Since 2003 has grown from 18 to 31 contributing archives with combined records increasing from 28,775 to 55,371. "Lessons learned" put into practice in new Emory University and DLF projects (see Aquifer affiliated projects in this report, including The Southern Digital Archives Conspectus from Emory).

- Documents from Emory's MetaScholar Initiative, including reports about the MetaArchive Project, AmericanSouth, MetaCombine, Music of the Social Change, and proceedings from the Workshop on Applications of Metadata Harvesting in Scholarly Portals: Findings from the MetaScholar Projects: AmericanSouth and MetaArchive (Halbert 2003) are available from
  http://www.metascholar.org/documents.html.
- AmericanSouth Final Project Report submitted to The Andrew W. Mellon Foundation, March 1, 2004: http://www.metascholar.org/pdfs/AmSouth-FINAL-REPORT.pdf
- Used as basis for advanced visualization and semantic clustering R&D in the MetaCombine project.
  http://www.metacombine.org/
- MetaCombine Project Interim Report on experiments in combined searching, automated organization of OAI and Web resources, and new forms of scholarly communication, September 2004.
  http://www.metacombine.org/reports/project/
- "SouthComb" Project will continue to develop the content, technologies, and practices developed in AmericanSourth.org. (Refer to section 4.4.8)

Archon

http://archon.cs.odu.edu/

Goal: To build a DL that federates Physics collections with varying degrees of metadata richness.

Status: No survey response received.

Most recent presentation available at Web site from 2003 covers Archon's architecture, equation and formulae searching, reference linking, conclusions and future directions. It announces plans:

- to make a separate service that can be plugged in or out which will make development changes easier and allow third party development of similar services
- to modify reference resolution to be able to refer to documents whether inside or outside the digital library
- to make this service context dependent using some of the technologies such as OpenURL to make the resolution dependent on the location and privileges of the user
- to work on the secure management of user accounts and access privileges using Shibboleth architecture to enable different access levels depending on use groups such as students or faculty (Digital Library Group, Old Dominion University, ICNEE PPT, available from http://archon.cs.odu.edu/publications.html .

Archon is also discussed in "Lessons Learned with Arc, an OAI-PMH Service Provider" (Liu et al. 2005)

ARL Scholars Portal Project
http://www.arl.org/access/scholarsportal/
Goal: To provide software tools for an academic community to have a single point of access on the Web to find high-quality information resources and, to the greatest extent possible, to deliver the information and related services directly to the user's desktop.
Status: As of May 2004, the project became "self-managing" among participating institutions with functional implementations using Fretwell-Downing software (under 3-year contract) with campus-wide releases at Arizona State University, University of Arizona, Iowa State University, and the University of Utah, and limited releases at University of Southern California (USC), University of California – San Diego (UCSD), and Dartmouth College.
- Scholars Portal Project Report, issued by ARL, May 10, 2004
    http://www.arl.org/access/scholarsportal/SPupdateMay04.html
- See also "The Current State of Portal Applications in ARL Libraries" (Jackson 2004)
    http://www.arl.org/access/portal/PAWGfinalrpt.pdf

Cyclades
http://www.ercim.org/cyclades/
Goal: To develop an open collaborative virtual archive service environment supporting both single scholars as well as scholarly communities.
Status: CYCLADES was a R&D project in operation from 1 November 2000 to 31 August 2003,
supported by the IST Programme of the European Commission (project no. IST-2000-25456) and set up under the framework of the DELOS Network of Excellence on Digital Libraries. Refer to:
- "An Open Collaborative Virtual Archive Environment" (Straccia and Thanos 2004)
- "A Personalized Collaborative Digital Library Environment: A Model and an Application" (Renda and Straccia 2005)

ENC Online: Eisenhower National Clearinghouse for Mathematics and Science Education
http://www.goenc.org/
Goal: Online math and science K-12 resource center.
Status: ENC resources were available free-of-charge until its funding from the U.S. Department of Education funding terminated. Effective October 1, 2005, goENC.com switched to a subscription service at a cost of $349 per school.

Among the ENC projects discussed in the 2003 DLF report, the only post-ENC project still active is the NSDL Middle School Portal, (although previous relics from ENC still show in up in NSDL search results).

- NSDL Middle School Portal with Pathways to Mathematics, Science and Technology is available from
  http://msteacher.org/

Flashpoint (Los Alamos National Laboratory)

http://lib-www.lanl.gov/lww/flashpoint.htm

Goal: Multi-database search tool for internal use at LANL.

Status: No survey response received. About LANL Research Library's new digital object repository architecture refer to: "aDORe: a Modular, Standards-based Digital Object Repository
" (Van de Sompel et al. 2005).

UIUC Digital Gateway to Cultural Heritage Materials

http://oai.grainger.uiuc.edu/

Goal: OAI aggregator of cultural heritage metadata.

Status: Since 2003 number of collections grew from 25 to 31; records increased from 413,563 to 538,485. However, project is now static and under review for probable discontinuation in light of the new IMLS Digital Collections and Content project, discussed more fully in Section 4.4.2 of this report. "Lessons learned" put into practice in other UIUC & DLF projects.

- Publications and presentations from UIUC's OAI metadata harvesting projects available from http://oai.grainger.uiuc.edu/presentations.htm
- See also, "Current Developments and Future Trends for the OAI Protocol for Metadata Harvesting" (Shreeves et al. 2005)
- IMLS Digital Collections and Content Project
  http://imlsdcc.grainger.uiuc.edu/

XTCat

http://alcme.oclc.org/ndltd/SearchbySru.html

Goal: One-time extract of 4.3 million bibliographic records from OCLC's WorldCat.

Status: Dormant. Arc continues to aggregate the entire set of records. OAIster continues to aggregate the 8,255 records that point to full text. The NDLTD Union Catalog includes a regular harvest of ETD metadata from OCLC's WorldCat, as discussed in Section 4.2.6.

Appendix 4
Comparison of Top Twenty OAIster and ROAR Archives based on Record Count (March 11, 2006)

RESOURCE Ranked by Size	OAIster Top Twenty	ROAR Top Twenty
PictureAustralia • National image aggregation • Australia • http://www.pictureaustralia.org/	832,506	Not included.
CiteSeer Scientific Literature Digital Repository • Aggregation of computer & information science documents with citation analysis. • United States • http://citeseer.ist.psu.edu/	716,772	703,654
PubMed Central • Digital archive of life sciences journal literature. • United States • http://www.pubmedcentral.gov/	463,229	492,069
Citebase • Semi-autonomous citation index that harvests scientific e-prints. • United Kingdom • http://citebase.eprints.org/	416,464	Not included.
arXiv • Physics e-print archive • United States • http://www.arXiv.org/	329,544	356,199
Pangaea: Publishing Network for Geoscientific and Environmental Data • Archiving, publishing and distributing georeferenced datasets • Germany • http://www.pangaea.de/	316,229	Not included.
University of Michigan, Digital Library Production Service Collection • Collection of texts and images primarily in the humanities. • United States • http://www.umdl.umich.edu/	281,072	See Humanities Text Initiative below.
Library of Congress Digitized Historical Collections • Digitized manuscripts, photographs,	231,413	231,654

rare books, maps, sound recordings and moving pictures. • United States • http://memory.loc.gov/		
Networked Digital Library of Theses & Dissertations (NDLTD) OAI Union Catalog • Aggregation of e-theses and dissertations • United States • http://alcme.oclc.org/ndltd/	Only harvests 8,255 records will full-text representation from OCLC's XTCat. http://alcme.oclc.org/xtcat/. Harvests ETDs from other sources.	217,520
Institute of Physics • Learned society aggregation of physics' journals. • United Kingdom • http://www.iop.org/	208,272	Not included; journals are restricted access.
State Library of Victoria OAI Repository • Digitized collection of manuscripts, photographs, and maps. • Australia • http://statelibrary.vic.gov.au/	201,934	Not included.
California Digital Library • Aggregated digital resources held in libraries, museums, archives, and other institutions across California • United States • http://www.cdlib.org/	174,097	Includes 10,900 records from the University of California eScholarship repository.
Digital Academic Repository of the Universiteit van Amsterdam (UvA-DARE) • Institutional Repository • The Netherlands • http://dare.uva.nl/	Includes only 2,145 records with full-text representation.	137,805
CITIDEL: Computing and Information Technology Interactive Digital Educational Library • Consortial digital library for computer and information technology education. • United States • http://www.citidel.org/	136,693	Not included.
RePEc: Research Papers in Economics • International database of working papers, articles, & book chapters in economics in collaboration with the American Economics Association. • United Kingdom	132,452	51,611

• http://www.repec.org/		
OSTI OAI Repository (U.S. Department of Energy, Office of Scientific & Technical Information) • Institutional Repository • United States • http://www.osti.gov/	122,753	Not included.
Wageningen Yield (WaY) • Institutional Repository • The Netherlands • http://library.wur.nl/way/	Includes 5,938 records only with full-text representation.	119,846
Instituto Tecnologico y de Estudios Superiores de Occidente: Acervo General de la biblioteca • Institutional Repository • Mexico • http://www.biblio.iteso.mx/biblioteca/	No links to full text.	119,606
National Library of Australia Digital Object Repository • Collections of manuscripts, sheet music, images and maps • Australia • http://www.nla.gov.au/digicoll/oai/	108,517	Not included.
University of Southern California Digital Archive • Aggregation of photographs, maps, manuscripts, records, texts and sound recordings owned by USC and collaborating institutions. • United States • http://digarc.usc.edu:8089/cispubsearch/	106,546	Not included.
Capturing Electronic Publications (CEP) Archive (University of Illinois) • Web site harvesting and archiving system focusing on state documents currently used by Illinois, Alaska, Arizona, Montana, North Carolina, Utah, and Wisconsin • United States • http://www.isrl.uiuc.edu/pep/	98,238 (PDFs only)	Not included.
The Wolfram Functions Site • Collection of formulas and graphs of mathematical functions. • United States • http://functions.wolfram.com/	87,618	Not included.
National Institute of Informatics Metadata Database	79,702	82,229

• Aggregation of university scientific information resources. • Japan • http://www.nii.ac.jp/index.shtml.en		
CERN Document Server • Aggregation of e-prints, articles, photographs and other formats related to physics and Institutional Repository. • Switzerland • http://cdsweb.cern.ch/	Harvests 38,459 CERN-specific materials only with full-text representation.	74,986
DIALNET: Servicio de Alertas y Hemeroteca Virtual de Sumarios de Revistas Cientificas Espanolas • Alerting service with access to the contents of Hispanic scientific literature. • Spain • http://dialnet.unirioja.es/	60,979	63,352
Humanities Text Initiative, University of Michigan • Digitized collection of e-texts in the humanities. • United States • http://www.hti.umich.edu/	See University of Michigan above.	58,632
CCLRC ePublication Archive (Council for the Central Laboratory of the Research Councils) • Central repository of publications containing the scientific and technical output of CCLRC. • United Kingdom • http://epubs.cclrc.ac.uk/	Harvests 13,599 records only with full-text representation.	55,511
SciELO: Public Health • Online health science articles with integrated access for Ibero-American countries. • Brazil • http://www.scielosp.org/	61,948	51,032
HAL: Hyper Article en Ligne • E-print server in France for all disciplines. • France • http://hal.ccsd.cnrs.fr/	56,140	51,692
University of Twente Repository • Institutional Repository • Netherlands • http://doc.utwente.nl:9080/en/	1,526	43,758
Demetrius Australia National University	42,601	42,602

Institutional Repository • Institutional Repository • Australia • http://dspace.anu.edu.au/		
Max Planck Society Edoc Server • Aggregated research output of the institutes of the Max Planck Society. • Germany • http://edoc.mpg.de/	Excluded because most identifiers do not link to digital objects but rather invoke SFX to find full text available through various proprietary online services.	39,944
Gallica, bibliothèque numérique de la Bibliothèque nationale de France (National Library of France) • Digitized collections, images and sound recordings. • France • http://gallica.bnf.fr/	36,390	36,398