Libri, 2005, vol. 55, pp. 181–197
Printed in Germany · All rights reserved

Libri
ISSN 0024-2667

Digital Repositories: Not Quite at Your Fingertips

NANCY JOHN
University of Illinois at Chicago, United States

The digital repository is a key technology used by today's libraries to collect, organize, archive and make accessible electronic files of different types. This paper argues that while the vision of the role of the digital repository has grown sharper and more articulate, the actual practical outcome has not met the hyperbole. Building blocks continue to be developed, but user access to repositories is still in its early development. There are promising exemplars of this technology, but more effort is needed. Particularly promising is some vendor open source work that may provide the tools needed to open up these digital resources. But fundamental change in how the existence of these repositories and their content is made known to the online user community is needed; traditional metadata access and harvesting is not enough. Infusing the content with an information context may be one way to assure that repositories are a significant part not only of the library of the future but also of the world's information landscape.

Introduction

Since the appearance of the first computer, more than a half century ago (CNN 1996), librarians and archivists have struggled with the question of how to share and promote access to computer files and how to assure their long-term survival, however long 'long-term' might realistically be. Despite concerns arising from these discussions, little concrete activity occurred during the first forty years – what one might name The Worry Stage.

The Worry Stage was typified by the sharing of horror stories, mostly anecdotal (Rothenburg 1995, 42) and frequently related to information in the public trust, e.g. U.S. Census, U.S. National Aeronautics and Space Administration. How much data was actually lost, or accessible only at unacceptably high costs, is not known. But the result was a heightened interest in protecting, migrating (moving from one media or one technology to another) and assuring access to information.

Interestingly enough at the same time that concern about access to electronic information became widespread, concerns about the availability of paper-based information also emerged. Worries about the loss of information printed on acidic paper were initially voiced as concerns about the estimated "80 million books in the nation's research libraries [that] were in danger of disintegrating because they were printed on cheap, wood-pulp paper, which many publishers used between 1840 and 1980" (Marcum 2002). Efforts in the 1980s and 1990s helped address this problem and correspondingly we felt a level of comfort that "we were doing something" to mitigate this situation.

Ultimately paper information was generally considered 'safe', given the use of acid-free paper, the storage of the information in multiple copies in libraries around the world and the conversion of acidic items to other formats. Marcum and Kenney (2002) noted that the research libraries addressing the brittle books problem were in the business of saving information for the long-term (archiving), but not in the business of making these resources widely available. The issue of wide access to these materials was a critical question

This article is based on a chapter in a work (in progress) about digital library technologies entitled *Not Quite at Your Fingertips*.

Nancy John is Digital Publishing Librarian and Associate Professor (retired), University of Illinois at Chicago, PO Box 8198. Chicago IL 60680. Tel: +1 312 996 2716. E-mail: nrj@uic.edu.

given the traditional role of research libraries to serve their own constituents. But open access was considered anathema to the long-term survival of paper information. Survival depended upon the locking up of some copies for safekeeping, while sacrificing other copies for access.

Another villain was attacking long-term availability of information – money. Long-term survival of the physical book or journal wasn't enough; willingness of some research libraries to share their resources beyond college walls wasn't enough, if no one could afford to acquire information, or if the terms of acquisition required significantly reduced accessibility.

The Scholarly Journal Crisis

The world's scholarly discoveries were being recorded for posterity in the thousands of specialized journals that had developed during the latter half of the twentieth century. The post-war explosion in scientific knowledge that began in the 1950s and 1960s, and continues today, resulted in hundreds of new academic specialties. As Abelson wrote in 1987, "One of the stimuli for scholarly publication is the belief by scientists and other authors that their work will add enduring values to the human heritage." Each specialty journal was fueled by the need to share research findings in the field, the pressure for scholars to publish (not perish) yielding many manuscripts, and a willing library to purchase the item to fuel both academic egos and more articles. And the perfect cycle was born – a captive set of writers and a captive set of readers and buyers.

These specialty journals collectively and individually held, and to a certain extent, despite some successful efforts to open up access to back files, still hold, a monopoly ownership of important discoveries in science, medicine, technology and their subfields. The original noble idea that sharing this information would fuel innovation and creativity, changed arguably to a notion that there was 'gold in them there hills' of scholarship. The staggering and increasing prices that could be charged in the marketplace, primarily to research libraries, had the concomitant, but unintended effect, of reducing access to information in the humanities and social sciences as libraries struggled to maintain high-priced subscriptions to these journals.

A third related concern was the issue of how the future generation would assimilate all this information and knowledge. In a landmark work published in 1989, Cetron and Davies asserted

> by the time today's [i.e. 1989's] kindergartners graduate from high school, the amount of knowledge in the world will have doubled four times. The Class of 2000 will be exposed to more information in one year than their grandparents encountered in their entire lives. They will have to assimilate more inventions and more information than have appeared in the last 150 years. (Cetron and Davies 1989, 65)

Twelve years later, the press would be discussing the problems that can occur when information is not easily accessible online, especially when older, key medical findings appear only in paper and may be overlooked by health care workers (Hopkins 2001).

Most of this is well documented in the published record, and not much can be said to add to the sorry state of affairs. But like every other period of stress, out of adversity, can come some creative solutions, or at least, the promise of some solutions. The need for more online-accessible information, the permanence of which is assured, at a reasonable cost that would keep libraries in the business of archiving and providing access to the world's assembled knowledge were key factors that led to the development of the information technology management strategy named "the digital repository," and its cousin "the institutional digital repository."

Because universities and other research institutions were seemingly caught in the vicious cycle of creating much of the world's knowledge but then giving it away to publishers so these same publishers could sell back the information, a special sub-type, dubbed the "institutional repository" appeared; the institutional repository referred specifically to the responsibility of institutions, such as universities or other research organizations, to gather together the knowledge created by their own researchers and to assure the long-term affordable access of this knowledge.

What is a repository?

The Oxford English Dictionary defines 'repository' as "a vessel, receptacle, chamber, etc., in which things are or may be placed, deposited, or

stored," and traces its first use back to 1485 in William Caxton's 'The Lyf of Charles the Grete'. In the mid-twentieth century, 'repository' had been used to signify a collection of items including documents and other types of objects; typically these materials were primary resource materials, i.e. original documents, not copies or secondary research materials. If we turn to the Web to understand what a repository is, Google retrieves several dozen definitions via a "define: repository" search (see Table 1 for a selection). A review of these definitions and the use of the term 'digital repository' (see Table 2 for the two definitions retrieved by Google) in the current research literature leads to the conclusion that the term signifies a collection of digital objects 1) accessible by using some sort of protocol(s) promoting their discovery, display and use by computer programs or by users directly and 2) maintained to promote long-term accessibility.

Table 1. Selected results from Google of the search, define: repository, accessed July 1, 2005.
Definitions of repository on the Web:

wordnet.princeton.edu/perl/webwn	depository: a facility where things can be deposited for storage or safekeeping a person to whom a secret is entrusted a burial vault (usually for some famous person)
en.wikipedia.org/wiki/Repository	A repository is a central place where data is stored and maintained. A repository can be a place where multiple databases or files are located for distribution over a network, or a repository can be a location that is directly accessible to the user without having to travel across a network.
www.orafaq.com/glossary/faqglosr.htm	A facility for storing descriptions and behaviors of objects in an enterprise, including requirements, policies, processes, data, software libraries, projects, platforms and personnel, with the potential of supporting both software development and operations management. A single point of definition for all system resources.
gams.nist.gov/Glossary.html	A repository is an Internet site that maintains and distributes a collection of software packages.
www.georgetown.edu/uis/ia/dw/GLOSSARY0816.html	A mechanism for storing any information about the definition of a system at any point in its life-cycle. Repository services would typically be provided for extensibility, recovery, integrity, naming standards, and a wide variety of other management functions.
memory.loc.gov/ammem/techdocs/repository/gengloss.html	A facility for storing and maintaining digital information in accessible form. A place where collections of digital information are stored. Also referred to as a "digital archive." In the NDLP context, digital information objects stored in the repository include materials such as sound recordings, text, pictures, photographs and moving images that have been converted to electronic form.

Table 2. Results from Google of the search, define: digital repository, accessed July 1, 2005.
Definitions of digital repository on the Web:

www.bl.uk/about/strategic/glossary.html	An organization that has responsibility for the long-term maintenance of digital resources, as well as for making them available to communities agreed on by the depositor and the repository. (www.rlg.org)
www.edtechpost.ca/pmwiki/pmwiki.php/Main/GlossaryAnalysis	A collection of digital assets and/or metadata accessible via a network without prior knowledge of the digital repository's structure. A repository is a network accessible server that can process the 6 OAI-PMH requests in the manner described in this document. A repository is managed by a data provider to expose metadata to harvesters. http://www.openarchives.org/OAI/openarchivesprotocol.html#Introduction

Key points repeated in these definitions are the accessibility of the objects and their long-term availability.

The Institutional Digital Repository

The institutional digital repository has been at the center of the library's responsibility to its parent organization. It certainly fulfils the library's role as the archive and safe-keeper of corporate information, but increasingly it is being seen as a way to extend the research and publication activities of the organization. Its electronic nature has however opened up the possibility of including all types of electronic files including for example email, datasets, and previous drafts. In a paper environment, one may well have expected written correspondence to be included in the archive of a scientists work, along with notebooks of research results, typescripts of manuscripts at various points in their development. Typically these became available at the end of a researcher's life, and access to them was limited to those who could identify the existence of the documents and travel to their whereabouts.

The development of the digital repository – an electronic archive – has changed this. First, electronic documents are notoriously ephemeral (Koehler 2004; National Archives 2004) and despite efforts to contain their destruction, much information is being lost, some of it likely to be important. Thus it is now widely held that libraries and archives need to be more proactive in securing the collections that they would eventually receive, to assure their existence and accessibility. This has led to the need for programs of continuous collection development of personal digital libraries and collections (Beagrie 2005). While Beagrie points out that individuals are striving to collect and maintain their electronic "bits" themselves, one can easily see that a more directed approach that collects, protects and assures access will serve future researchers far better.

The institutional digital repository provides a method for capturing and maintaining today's electronic detritus so that tomorrow's scholars can understand the thinking behind the published record. In addition, the institutional repository provides a way for an institution to capture the more polished electronic works – books, articles, dissertations, technical reports, etc. – and to guarantee access by the organization, researchers worldwide and the public to balance out the loss of this information to the private sector under restrictive licensing agreements.

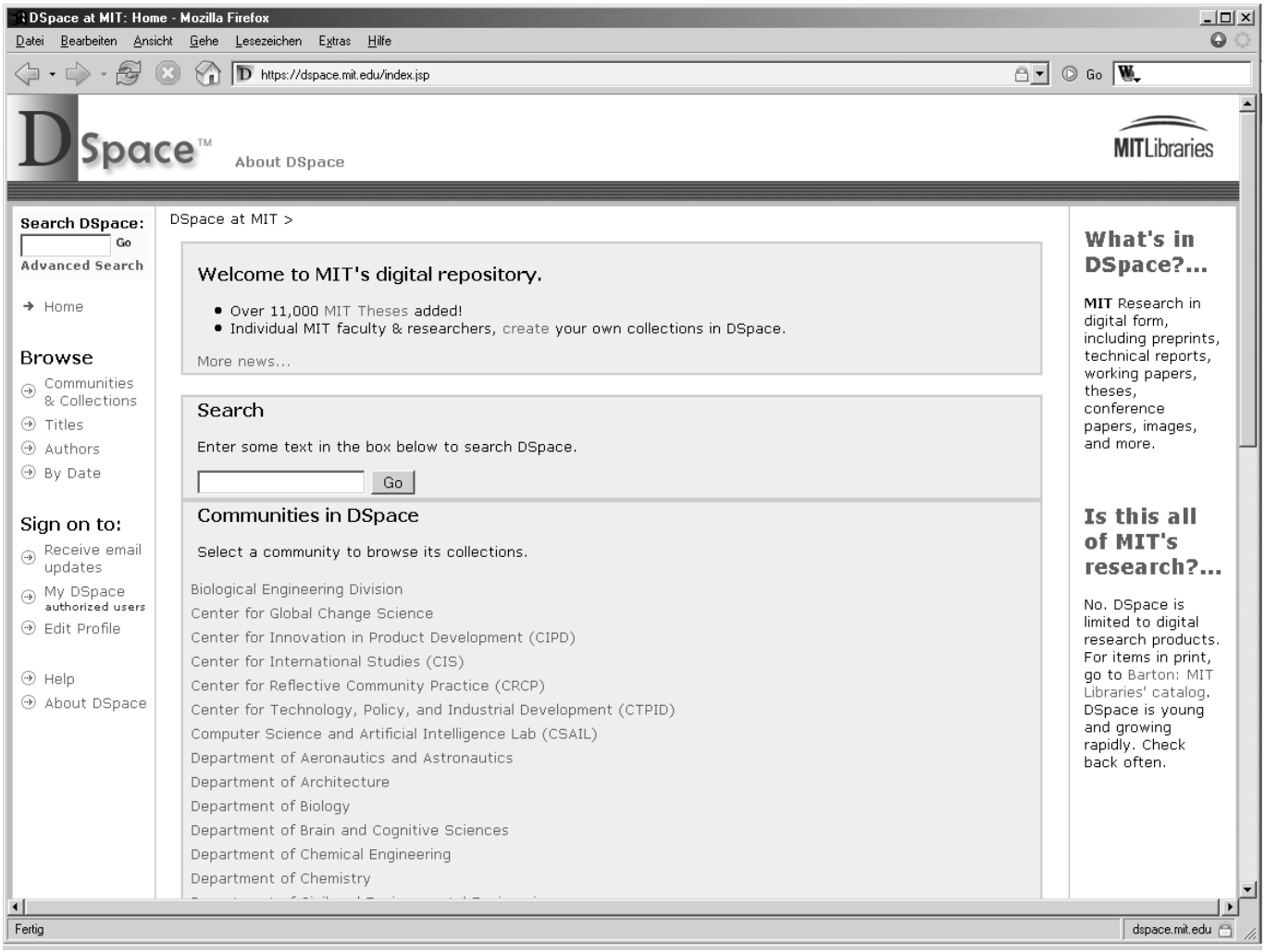
By providing a platform for an organization's researchers to showcase their intellectual achievements, the library is responding to a variety of concerns: long-term access, open access, and improved re-use of intellectual property. Initial reaction of researchers has been mixed. The rigors of the promotion and tenure process may lead more junior researchers, the ones who are more likely to support open digital access, to lock up their intellectual property in elite and expensive scholarly journals in order to improve their tenure evaluations. More senior authors may create more of their content in response to invitations that result in their output becoming unavailable to local repositories.

Despite the less than enthusiastic initial response, it was clear that much work needed to be done to create a reliable, attractive infrastructure. Through development of that infrastructure, it was hoped that the objections raised by scholars would eventually be mitigated. The need for technology to make repositories accessible and permanent has led to the development of a variety of software and protocols to assure interoperability and effective management of digital files.

Digital Repository Software

IBM issued its Digital Library software in 1991 as the first major effort at using automation and system architecture to manage collections of digital files. It initial release spurred some digital collection activity in the early 1990s. Notable projects include the Vatican project (Gladney 1997); Hermitage Museum, St. Petersburg, Russia; National Palace Museum, Taipei; and the original Variations project at Indiana University. IBM's groundbreaking technology grappled with key issues of storage, maintenance, retrieval and display of digital content. From these first efforts at building significant collections to open up access to collections around the world came a number of important software developments that moved the stewardship of digital content away from home-grown Web page exhibitions to sophisticated content management systems. The following is far from a complete catalog of the software options, but it identifies some of the major players in this arena.

Figure 1. DSpace at MIT.



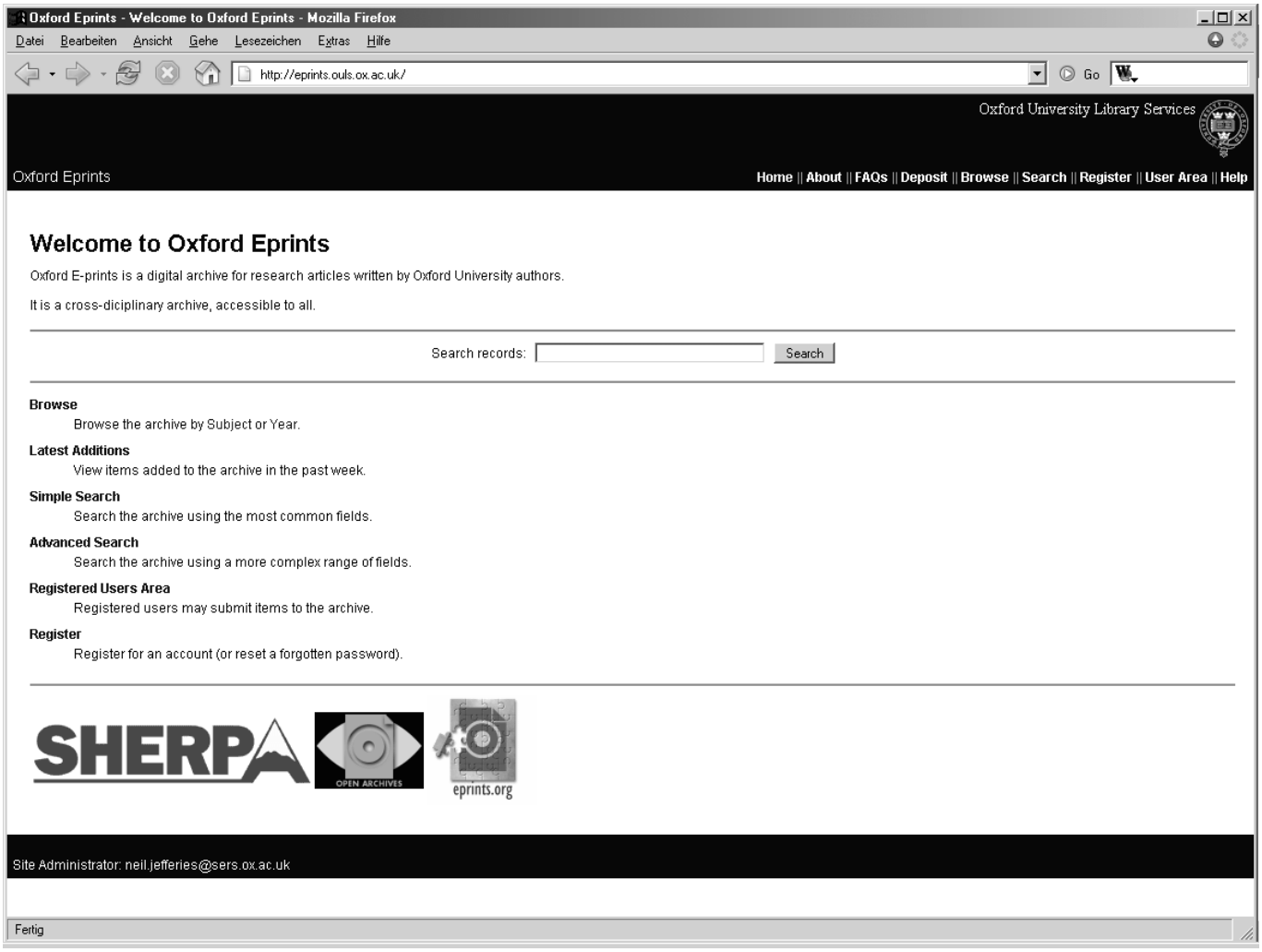
DSpace

DSpace (http://www.dspace.org) was developed jointly by The MIT Libraries and Hewlett-Packard (HP). DSpace modestly describes itself as "a groundbreaking digital repository system. DSpace captures, stores, indexes, preserves and redistributes an organization's research material in digital formats." Used worldwide to meet many digital archiving needs, DSpace supports institutional repositories, learning object repositories and electronic records management. The open-source system is available to research institutions worldwide as an open source system that can be customized and extended. DSpace has taken on the trappings of other software suites: users' group meetings and a federation to manage its development. The DSpace wiki provides considerable historical and technological information (http://wiki.dspace.org/).

DSpace is meant to manage institutionally produced research and teaching materials across disciplines. By recognizing the different communities within an institution DSpace can offer a federated approach to customized sub-repositories that reflect the particular values and needs of scholarly groups (see Figure1, for an example). The customized approach allows for appropriate review and access restrictions to be imposed. Retrieval is a simple Web browser transaction to locate and either display or download needed files. Qualified Dublin Core metadata assists in the location of files.

DSpace takes a cautious approach to long-term survival of the files. By dividing formats into a list of supported formats (those DSpace commits to

Figure 2. Oxford University's Oxford Eprints.



keeping 'alive' long-term), known formats (those DSpace hopes to keep 'alive' by partnering with the vendor community), and unsupported formats (those that DSpace either can't commit to long-term survival, or will only commit to survival by converting the files into another format), DSpace is managing the preservation of the functionality of the files. Using simple techniques such as checksums and file descriptions and signatures, DSpace also manages bit-level preservation.

The DSpace user community continues to grow; its growth beyond the United States promises to bring some interesting new approaches in the near future.

Eprints

Eprints (http://www.eprints.org) is the original digital repository software developed by the University of Southampton to manage an open archive. Released in 2001, it was the first Open Archives Initiative (OAI)-complaint repository software. As implemented, it typically supports collections of pre-prints and technical reports, often subject-based in scope. More recently several universities have implemented the software to manage multi-disciplinary institutional archives. It may be difficult for the casual observer to sense a difference in exploring the various public views of DSpace or Eprints repositories. But Eprints was developed from the standpoint of encouraging scholars to self-archive their work – and in that sense there is not a community review and decision process. Eprints's strength comes from the agglomeration of individuals' self-archiving actions. The authors of Eprints have a cause, opening up scholarship by encouraging authors to keep their copyrights or to assert themselves

by demanding permission to self-archive their work after publication. Research by the evangelists of open archives Eprints (Harnad and Brody) shows that materials deposited in open archives receive greater use (as measured by citation). The software is used worldwide including CalTech and for the SHERPA (Securing a Hybrid Environment for Research Preservation and Access) Project, which seeks to explore the impact of open access to research and is funded by JISC (The Joint Information Systems Committee) and CURL (The Consortium of University Research Libraries in the British Isles) and hosted by the University of Nottingham.

There are 195 known archives using EPrints worldwide (see Figure 2 for an example). The Eprints.org site also keeps track of which journals and publishers support self-archiving of preprints and/or published articles.

Fedora

Fedora (Flexible Extensible Digital Object and Repository Architecture) is a digital repository system developed jointly by Cornell University Information Science and the University of Virginia Library. The Fedora Project's goal is to provide open-source repository software and related services to serve as the foundation for many different types of information management systems. Fedora is not a complete system in the sense that DSpace and Eprints are – instead it provides an infrastructure upon which services can be developed. One can think about Fedora as the software that creates the library building, the bookstacks and the classification system, but then needs ancillary services – like a catalog – to complete the picture. But Fedora provides much more than this simple analogy because of its structure which not only requires but also promotes the building of customs tools to expose the repository in creative ways. So while the items in the repository can play together, front ends that make encyclopedias (http://www.encyclopedia.chicagohistory.org/), a statewide digital highway (http://www.njdigitalhighway.org/) and a library's digital collections (http://dl.tufts.edu/). One of the interesting challenges for Fedora users is the need to make tools to expose the data managed under Fedora. One effort to provide a more flexible open source solution to sit on top of Fedora, is the Fez project.

Fez (http://espace.library.uq.edu.au/) aims to provide a freely available and open-source Web-based Digital Repository, specifically for use in Libraries; its initial implementation showcases research documents and data at the University of Queensland Library in Australia. "Fez is an open source project to produce and maintain a highly flexible Web interface to FEDORA for any Library or Institution to configure and publish or archive documents of any type sustainably." (Fez n.d.) The site also lists a number of goals for the project, among which some of the most important are:

- peer review and research reporting interfaces
- automatic archival to Web-format datastream (file) conversion
- automatic object provenance and changes/history logging
- commenting system
- annotation system
- federated searching
- integration with a Handles server
- integration with a Google Web crawler service provider
- dynamic canned search links (e.g. for authors, controlled vocabularies)
- tie-in for creative commons licensing
- full text indexing/searching (e.g. for pdfs).
- conversion of pdf/Word files into xml format. (Fez n.d.)

Greenstone

Greenstone (http://www.greenstone.org) is software for building and distributing digital library collections, produced by the New Zealand Digital Library Project at the University of Waikato, and developed and distributed in cooperation with UNESCO and the Human Info NGO, and issued as open-source, multilingual software, under the GNU General Public License.

Examples include the Welsh Book Council's Books from the Past (http://www.booksfromthepast. org/), the Chopin Early Editions (http://chopin.lib.uchicago.edu/), Ulukau, Hawaiian Electronic Library (http://ulukau.org/) and the New Zealand Digital Library (http://www.sadl.uleth.ca/nz/cgi-bin/library), the original impetus for the development of the software (see Figure 3).

Figure 3. The New Zealand Digital Library.

Greenstone is particularly strong in its interoperability with other platforms and protocols.Not only can it serve and harvest documents and collections over the Open Archives Protocol for Metadata Harvesting (OAI-PMH) but also collections can be exported to or imported from METS (Metadata Encoding and Transmission Standard, www.loc.gov/standards/mets/), collections can be exported to DSpace via DSpace's batch import program, and DSpace collections can be imported into Greenstone.

Commercial alternatives

CONTENTdm®

CONTENTdm® (DiMeMa, distributed by OCLC) was developed at the University of Washington. Its developer created a spin-off company early in 2001 to support the growing user community and to focus on accelerated research and product development. Early research by the developers looked at how information was shared on the World Wide Web. This led to the development of a technology for sharing media collections on the Web for the University of Washington Libraries (see Figure 4) special collections that existed in a variety of forms and formats. Interest form the library community outside of the University led to the development of CONTENTdm® Digital Collection Management Software, now distributed by OCLC, Inc. The software has tools for acquiring or creating collections, tools for storage of the content and a set of tools for display and retrieval of objects. The University of Washington has more than 75 collections in CONTENTdm® (http://content.lib.washington.edu/).

Figure 4. University of Washington Libraries Digital Collections.

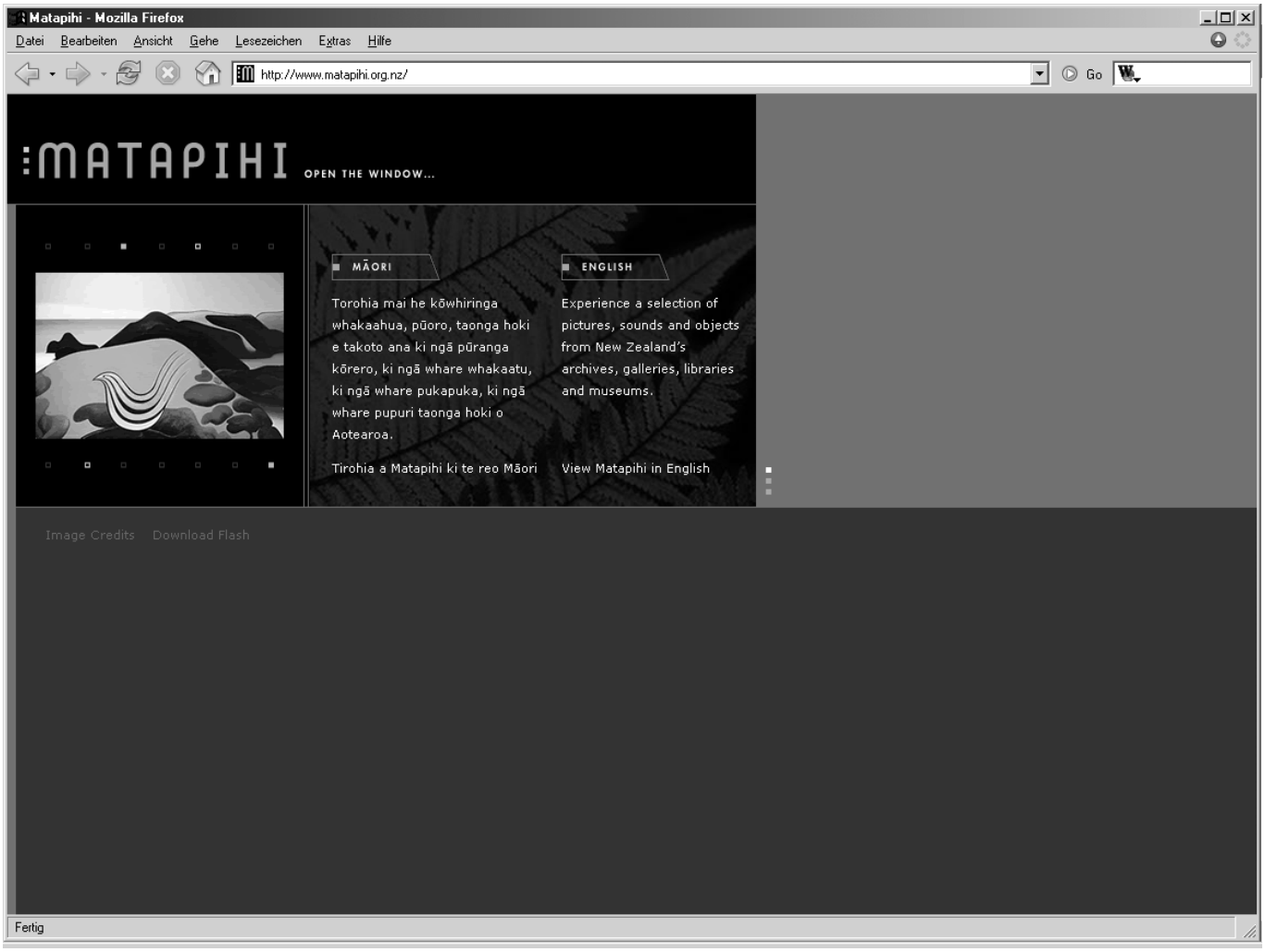
DigiTool

DigiTool (Ex Libris) is "an enterprise solution for the management of digital assets in libraries and academic environments." (Ex Libris n.d.) Institutions create, manage, preserve, and share locally administered digital collections that can be integrated with institutional portals and e-learning systems. DigiTool asserts that its modular architecture is designed to address the current and future requirements of a broad range of digital collection types including institutional repositories, collections of educational materials, and library and museum special collections. DigiTool has modules designed to address the workflows needed for the life cycle of a digital object: metadata, management of the objects and end users search. The DigiTool Repository uses a standard Web services (SOAP) layer to enable the repository to interact with the other DigiTool modules as well as with local or third party systems.

ENCompass

ENCompass for Digital Collections (Endeavor) is offered by Endeavor as part of its suite of software for managing and accessing digital content. The other modules include ENCompass for Resource Access, ENCompass for Journals OnSite (EJOS), and ENCompass Course Content Integrator. An excellent example using this software is the Matapihi site of New Zealand cultural heritage built by the National Library of New Zealand (see Figure 5). In addition to helping libraries to create and manage digital collections, ENCompass aims to help libraries unite collections previously made under different software and using

Figure 5. Matapihi



a variety of metadata standards under a single front-end user experience.

Hyperion

Hyperion (Sirsi) provides organization, storage, and access to digital files by searching both associated metadata and full-text of text files. It is based on the Sirsi automated system, and integrates with it. Rice University's Fondren Library maintains such a site at http://www.rice.edu/fondren/hyperion/.

MetaSource

MetaSource (Innovative Interfaces) is software for describing and digitizing media collections offering a way to store, crawl, index and describe these collections, and then to integrate them into the catalog or maintain them as separate searchable collections. MetaSource is a suite of tools used to manage digital collections, including digital object storage, crawling external collections and support for metadata schemes. MetaSource is made up of three components: Millennium Media Management, XML Harvester, and MetaData Builder. Millennium Media Management creates and stores media objects such as images, sound files, and video files. The XML Harvester gathers XML records from any server and inserts them into the local catalog. MetaData Builder stores XML metadata in the library's preferred scheme. There is a Copyright and Access sub-module to provide controlled access to digital collections with access restrictions, such as electronic reserve material. Also this unit also assists the library with the actual scanning (creation) of digital objects.

VITAL

VITAL (VTLS) is institutional repository software which is a set of workflow extensions, management utilities, and enhanced searching capabilities built on the Fedora repository architecture. It creates new tools and enhances the functionality provided by Fedora itself. Designed to simplify the development of digital object repositories and to provide seamless online search and retrieval of information, VITAL provides for ingesting, storing, indexing, cataloging, searching and retrieving collections. Using Web services, VITAL provides a mechanism to create tools, enhance the functionality provided by VTLS or leverage the open-source community for future applications. Among the key features it promises are:

- storage and management of any content format due to VITAL's repository architecture
- integration with existing systems through open, standards-based protocols
- search full-text content of PDF, DOC, RTF and other document formats
- display high resolution imagery, multi-page documents and specialized data formats
- automatically capture preservation metadata and create long-term, citable DOIs (VTLS n.d.)

On October 21, 2005, VTLS proved its commitment to the open-source community by announcing the availability of free, open-source components that will work with FEDORA™ and/or VITAL.

The open-source software, developed by VTLS, on behalf of ARROW (Australian Research Repositories Online to the World), consists of:

1. The Metadata Extraction Service via JHOVE (JSTOR/Harvard Object Validation Environment)
2. The Handles System for assigning, managing, and resolving persistent identifiers, known as "Handles" within the FEDORA™ repository.
3. The Content Model Configuration Service for customized content models to be created defining objects of similar content.
4. The SRW/SRU (Search/Retrieve Web and Search/Retrieve URL) Interface for exposure of repository content that defines a method for interacting with and retrieving information from remote databases and to expose the content of the FEDORA™ repository to portals, federated search tools, and other search engines which support this emerging protocol.
5. The Web Crawler Indexing and Exposure Service to expose repository content to Web crawlers, such as Google, via a MARCXML to XHTML conversion of metadata. [1}

Evaluating repository options and digital content managements software

With such a wide variety of software options, it's no wonder that digital repositories are only in their infancy. Each software developer hypes its own packages pointing out the flaws of the others. Fortunately, a few brave institutions are sharing their methods for evaluating software solutions. One such institution is the University of Arizona (Yan Han 2004) whose evaluation strategy is more useful than the particular findings.

The Budapest Open Archive Initiative provides the very useful "A Guide to Institutional Repository Software v 3" at the URL, http://www.soros.org/openaccess/software/. The guide covers Archimede, ARNO, CDSware, DSpace, Eprints, Fedora, i-Tor, MyCoRe and OPUS. Its "Feature & Functionality Table" is an excellent guide to evaluating any software for managing digital repositories. JISC also has a very good site describing the evaluations made by its FAIR (Focus on Access to Institutional Resources) participants (http://www.jisc.ac.uk/index.cfm?name=fairsynthesis_repsoft).

Digital Repository Protocols

Lynch (2003) argues quite articulately that there is much confusion about the application of the term "digital repository." He notes that it has come to denote both the actual collection of files and the suite of services that makes those files accessible. For the sake of understanding the state of the practice of digital repositories, this article has separated out some significant services that enhance the retrieval, long-term viability, processing and management of digital repositories.

Preservation

Preservation of digital objects is a subject worthy of far greater treatment than this brief summary. However, the longevity of digital objects is one of the most pressing problems facing today's digital librarians and researchers. One of the first agencies to look systematically at the survival of digi-

tal content was the National Library of Australia through its PADI (Preserving Access to Digital Information) project. It aims to provide mechanisms "to ensure that information in digital form is managed with appropriate consideration for preservation and future access. Its objectives are:

1. to facilitate the development of strategies and guidelines for the preservation of access to digital information;
2. to develop and maintain a Web site for information and promotion purposes;
3. to actively identify and promote relevant activities; and
4. to provide a forum for cross-sectoral cooperation on activities promoting the preservation of access to digital information." (National Library of Australia n.d.) This gateway to information about preservation is excellent. Bur still one can not help but wonder why, despite the growing number of links, there are so few highlights in solving the problem of the preservation of digital information. What are clearly needed are more experiments testing the various approaches to both the physical and the functional preservation of digital files.

One such experiment is LOCKSS (Lots of Copies Keep Stuff Safe). The LOCKSS technology was first proposed in 1999. A beta version was deployed to 50 libraries worldwide from 2000 to 2002. From 2002 through mid 2004, the Stanford University LOCKSS Program team developed the program that was released into production April 2004. The idea is simple; different libraries keep copies of the same electronic journals and from time to time compare their electronic cache's of the journals to assure that there are multiple complete and audited copies of publisher's electronic journal sites. This notion of multiple copies, spread around, to keep the world's knowledge safe for posterity, was first proposed by Thomas Jefferson.

But the multiple copies solution only addresses the existence of the files. And while one hopes the regular comparison of caches might address the question of whether the files are complete as created, there is still work to do to assure that files are exactly as created. Use of digital signature, checksums and file descriptions all help, but still more work is needed in the refreshing of files at the bit-level. And even if the file is complete, it may not be readable. The question of what techniques support the migration of files so they can be used indefinitely has not been answered. Until we've solved the problems of permanent storage media and permanently usable file formats in permanently readable storage, digital preservation is anyone's best guess. But until it is solved, digital repositories will be for near-term access and not long-term survival.

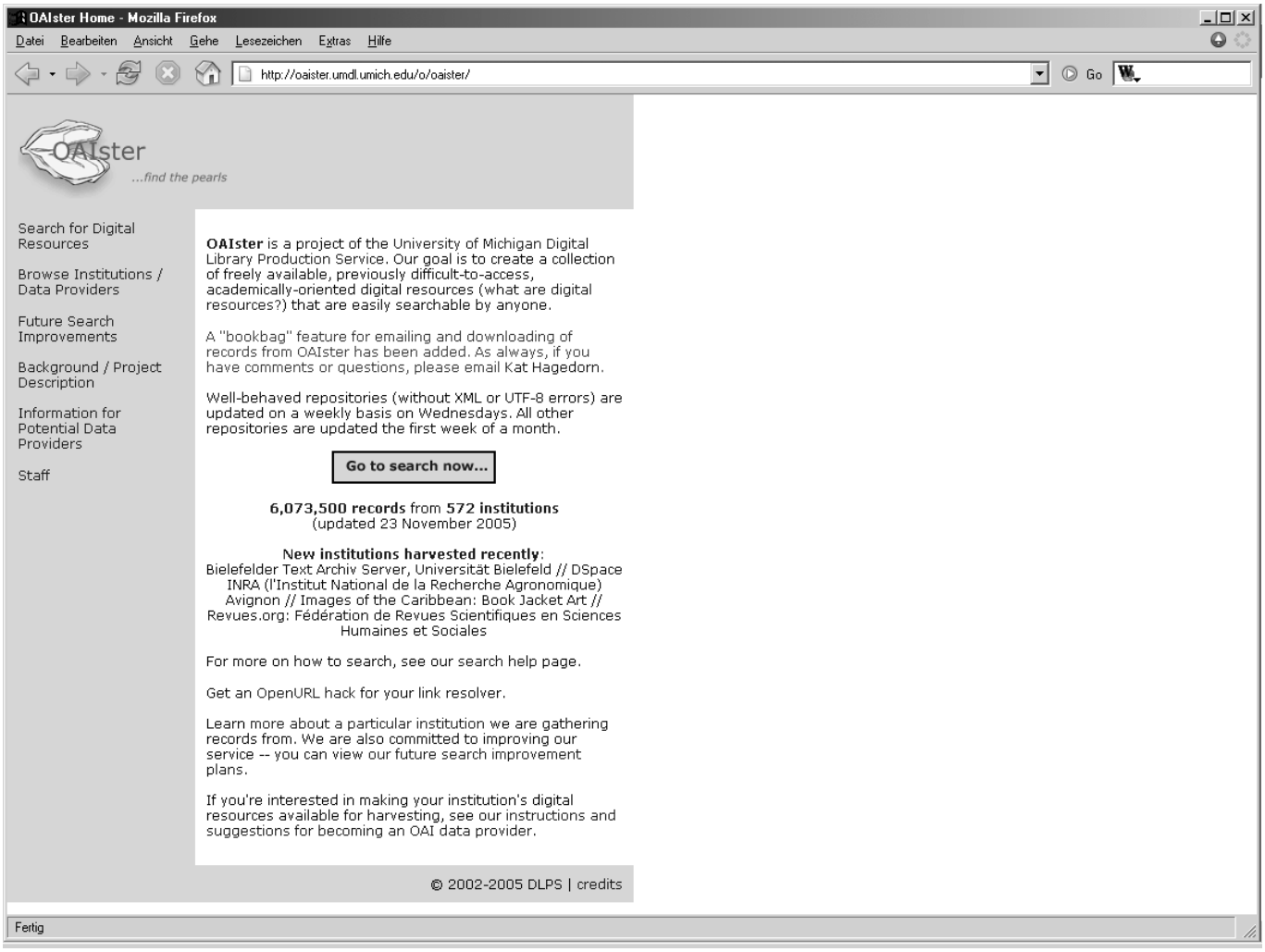
Searching and retrieval

The searching and retrieval of items in a digital repository is a two-fold problem. The first is the searching inside the repository, which is addressed by the repository software. The second is how to search across multiple archives. Google, Yahoo! And the other Web search engines have ably demonstrated the power of searching across millions of Web pages. Searching in a single repository can only be improved by searching across multiple repositories. Searching across an institution's repositories will be further enhanced by searching across the repositories of multiple institutions. A protocol that will assist with the discovery, searching and retrieval of digital objects held in digital repositories would be quite useful, or such was the thinking of the developers of the Open Archives Initiative – Protocol for Metadata Harvesting (OAI-PMH) (http://www.openarchives.org/OAI/openarchivesprotocol.html), now in version 2.0.

It provides an application-independent interoperable framework for metadata harvesting consisting of *data providers* that exposing metadata via OAI-PMH and *service providers* that use harvested metadata as the basis for building value-added services. As of September 2005, he Open Archives Initiative lists 359 repositories and 22 service providers. OAIster (http://oaister.umdl.umich.edu/o/oaister/) at the University of Michigan is one of the largest and most complete with more than 6,000,000 records from 572 institutions (as of November 2005). For a sample search, see Figure 6.

The concept of cataloging an item and aggregating the catalog records for a group of items is a time-tested method used by libraries to access their paper collection. But, as today's indexing and abstracting services have learned, aggregated metadata is not enough. High quality metadata is easily passed up for one click access to content. Services like Google, Yahoo! and Amazon.com have earned their places in the information landscape by adding information that puts retrieved

Figure 6. Searching OAIster for 'open access repositories'



data in context. With so many ways to access online information already available to the user, library digital repositories must compete for their fair share of search-time (i.e. fingertip time) if they are to be among the favorites and oft-visited sites of their users.

But before we look at how libraries might respond to these challenges, it is useful to look at some interesting examples from a content viewpoint to convince ourselves that improving access to the repositories is warranted. After all, easier finding only works if the information is there in the first place.

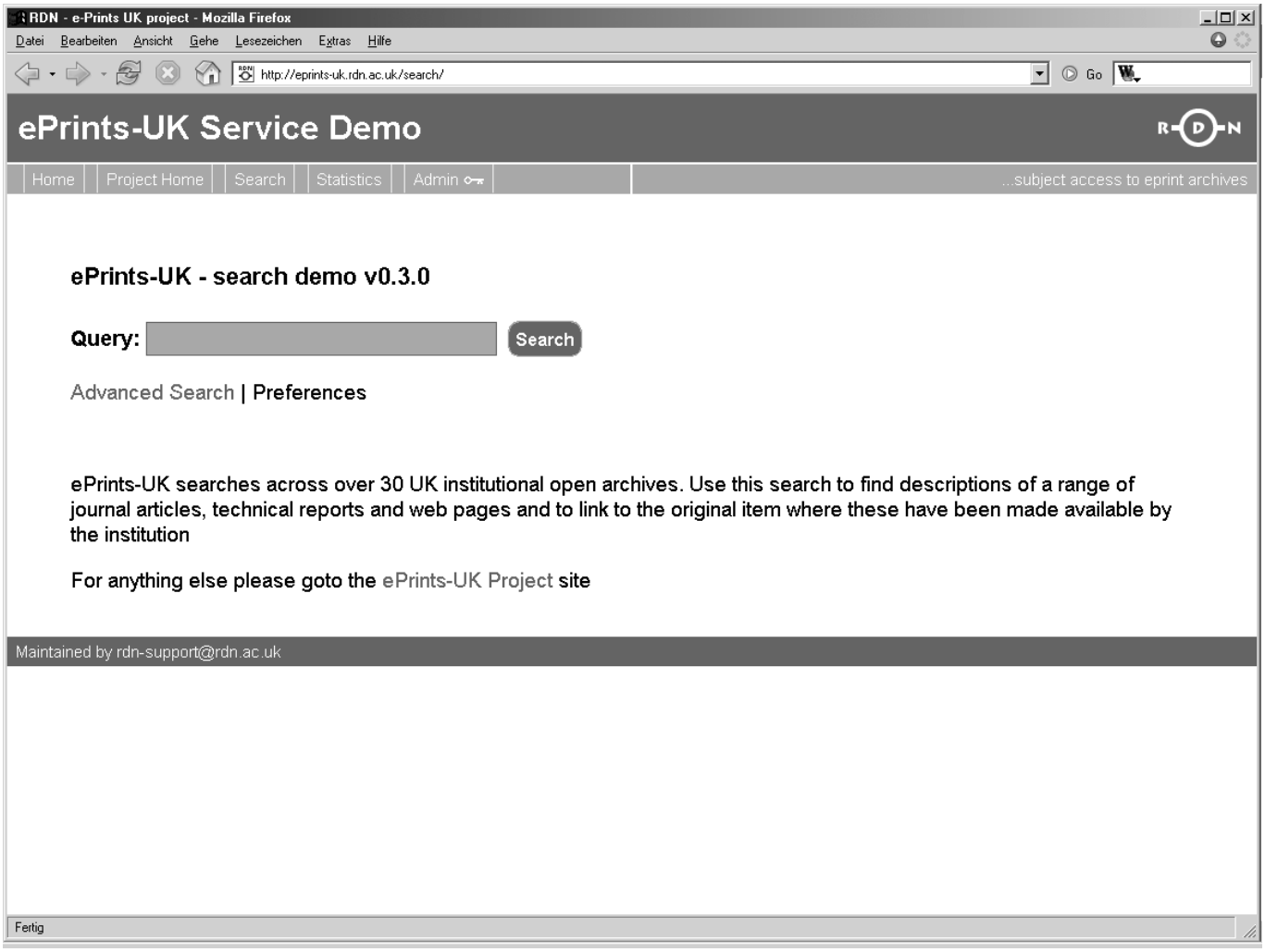
Some Interesting Examples

Librarians are not the only creators of digital repositories although they are clearly a significant force in promoting their development. Actions by professional societies are also having an impact on the creation and development of digital repositories. For example, a group of communications societies have created a new digital repository in communications,

> Academic Serials in Communication – Unified System (ASCUS, pronounced "ask us") is an innovative, not-for-profit, society-governed, full text database of academic publications serving the academic communication field. ASCUS heralds a new relationship between academics, societies, and publishers in which content is widely distributed at low cost. (ASCUS 2005)

Proof of the attractiveness of this proposal for a rich repository of subject content came when ASCUS announced November 5, 2005, that a major publisher in the communications field, Lawrence Erlbaum Associates, Inc., had joined their effort. This digital repository has a built in community of users through its society participants.

Figure 7. A search of 'open access repositories' in the ePrints UK server.



The ePrints UK Project (http://eprints-uk.rdn.ac.uk/) is developing a series of national, discipline-focused services through which the higher education community can access the collective output of UK universities and colleges.

A search of the phrase 'open access repositories' reveals 10 items (see Figure 7).

Once again the collective activity from the outset may mean a richer set of content and content tools, a larger community of users and possibly a more successful experiment.

Library and Information Science Repositories

Given the important role played by librarians in developing digital repositories, one would expect that digital repositories in library and information science would be common. The reality is somewhat different although there are indeed some interesting projects.

ALIA e-prints

The Australian Library and Information Science e-prints server has papers from the Australian library community. The most recent year boast 39 conference papers (http://e-prints.alia.org.au/view/year/2005.html).

ArchiveSIC

ArchiveSIC – Archive Ouverte en Sciences de l'Information et de la Communication (http://archivesic.ccsd.cnrs.fr) is a self-archiving server for articles and working papers in the fields of the Information Sciences and Communication

(SIC) using the Eprints software. In November 2005, there were 637 documents on the server.

colLib

colLib (http://collib.info/index.php/Main_Page) is a collaborative platform for organizing Open Access materials in Library & Information Science (LIS) that is being developed and maintained by a graduate student in Tromsø, Norway to explore the concept of 'overlay' in metadata harvesting. "colLib harvests metadata-records from OAI-PMH-compliant repositories and enables manual 'tagging' of these records to cluster them by subject or other meaningful categories. Tags are represented by pages in a wiki, that can be annotated with links to related tags, external links and any other text deemed relevant." In September 2005, the total number of records in colLib was 4003. According to the colLib blog, "3469 (87%) of these have not yet been assigned any tags."

The Directory of Open Access Journals

The Directory of Open Access Journals (DOAJ) (http://www.doaj.org) lists 54 journals under the subject area of library and information science. Eleven of these journals have content that has been harvested and made available through DOAJ.

DLIST

DLIST (http://dlist.sir.arizona.edu/) is an open repository in library and information science developed using the OAI-PMH by the University of Arizona. Authors are encouraged to deposit their papers in the DLIST repository. The repository currently contains about 550 papers.

DoIS

DoIS: Documents in Information Science (http://wotan.liu.edu/dois/ and http://www.dois.it/) is a database of articles and conference proceedings published in electronic format in the area of Library and Information Science. It is a sister project to the E-LIS project. The most recent figures cite holdings of 13,403 articles and 4313 conference proceedings, with 12,236 downloadable from the site.

E-LIS

E-LIS (E-prints in Library and Information Science) is an open access archive for scientific or technical documents, published or unpublished, on librarianship, information science and related areas. It "relies on the voluntary work of individuals from a wide range of backgrounds and is non-commercial. It is not a funded project of an organization. It is community-owned and community-driven. We serve LIS researchers by facilitating their self-archiving, ensuring the long-term preservation of their documents and by providing word-wide easy access to their papers." (E-LIS home page 2005). It uses volunteer editors from around the world to advertise and promote its existence, and to manage the archive. As of November 2005, E-LIS had more than 3,000 pre-prints and post-prints from more than 200 journals.

e-Prints Soton

The e-Prints Soton (University of Southampton) has a subject area 665 (Library and Information Science) that lists 25 papers (http://eprints.soton.ac.uk/view/subjects/Z665.html).

METALIS

METALIS (http://metalis.cilea.it/) is a Service Provider collecting metadata from institutions that offer full-text papers and documents in the field of library and information Science.

While this brief survey shows that there is a number of promising repositories for library and information science information, it also shows that neither the number of repositories nor the amount of documents is sufficient to entice the average Web searcher into expending the effort needed to locate and search for them. It is also difficult to know how complete a survey of LIS repositories this is, and additional research is needed to assess how much content not normally available to a researcher can be located by using these sources. Projects like OAIster promise a single front end to searching but do not yet provide the focus that a subject repository can. However, as we librarians use these repositories more and more for our own research, we will see the limitations that other researchers experience and

will improve both their usefulness and attractiveness.

The Future of Digital Repositories

This brief survey is adequate to justify the statement that digital repositories have a future. But what kind of future? Will they remain interesting detours for scholars or will they become mainstream Web resources highly visited and highly ranked? For those who doubt the ability of the Web to survive because of its cacophony of information, there may be one answer; for those who believe that any content beyond an easy Google search will always be underutilized, there is another answer.

Just as library catalogs have led to technologies and standards that are used far beyond the catalog, digital repositories will no doubt improve the permanence and retrievability of all digital objects, whether in a digital repository or not. One thing is certain, despite the protocols and technologies, original unscientific, non-standard Web-based exhibitions published by the world's libraries remain far more accessible to the average user than the content-rich digital repositories they are creating. This is because they are part of the indexed, spidered, crawled open Web that is accessed via search engines. Digital repositories, for the time being, belong to the vast hidden Web, found through published citations, links and personal recommendations. For researchers, the ease of reviewing a remote collection through its digital surrogate can mean the end of trips to closed archives and long waits while items are located. Not that surrogates will replace originals, but if a picture is worth a thousand words, a digital object certainly exceeds the meager descriptions in library and archive typescript finding aids.

But there are differences between the library catalog and Amazon.com. While librarians may argue that library catalogs are more rigorously developed, users argue that Amazon.com is easier to use, more complete and provides additional information that helps the user to evaluate a title without actually holding it in one's hands (e.g. reviews, other books looked at by other searchers). Unfortunately there are few differences between the digital repository and its antecedent finding aid. Today's repositories are physically more accessible than the finding aid that was a filing cabinet or a pile on the corner of a scholar's desk, but a digital representation and some indexing metadata does little to expand the intellectual access. Even full-text indexing, unless it makes use of conceptual strategies, does more to clutter the information landscape than to organize it. Little effort has been expended on those additional features that could make a digital service really useful – context and relationship to other information.

The World Wide Web community collaborated to produce the 837,161 articles of the English Wikipedia in just four years, and the Wikipedia regularly rivals old warhorses like the Britannica in currency and completeness. Digital repositories are predictably staid despite the prevalence of wikis and blogs in academia. The use of these technologies could help to put digital resources into context and to lead the users of these collections to other likely sources, while attracting new users (or making for repeat users). The value of a running commentary on articles provided by previous readers/searchers would be highly prized by all researchers. Search histories of individual users could be exposed, while protecting the individuals' privacy, to enhance the relation of one article or paper to another. For example, a service might provide the following: "people who looked at this article also looked at the following 3 articles". Or features like Amazon.com's "Customers who searched for <search argument> ultimately chose:", ratings and excerpts could assist the searcher in selection of an item. Other Web information services provide similar features; for example, the Internet DVD-rental service Netflix (http://www.netflix.com) allows users to review movies to enhance the already linked reviews, to recommend movies to others and it makes recommendations based on items the client has looked at or added to her queue. The citation analysis pioneered by Eugene Garfield and most recently used to inform searchers of Google scholar only keeps track of used articles. Our technology today allows us to see how researchers chose not to use a particular article by coupling citation evidence with searching history. Valuable pathfinders to topics could be developed by revealing, in the aggregate, the search paths taken by scholars.

Huwe (2005) has three wishes for digital repositories: the use of creative Web pages to make

a whole out of the 'archipelagos' that are our digital collections; use dynamically generated pages to provide context information for the user; and finally, while there is evidence of activity, there needs to be more. His point is well taken that sometimes we librarians can't see the forest through the trees. When it comes to digital repositories, building and linking them is only just a beginning.

Notes

1. Excerpted from the October 21, 2005, VTLS Press release (http://www.vtls.com/Corporate/Releases/2005/24.shtml) that states "All of these functional components are available for download at the following link: http://www.vtls.com/Products/osc.shtml."

References

ASCUS. 2005. Home page. URL: http://www.ascus.info/ [viewed Nov. 10, 2005]

Beagrie, Neil. 2005. Plenty of room at the bottom? Personal digital libraries and collections. *D-Lib Magazine* 11(6, June). URL: http://www.dlib.org/dlib/june05/beagrie/06beagrie.html [viewed October 15, 2005]

Cetron, Marvin J and Owen Davies. 1989. *American renaissance: our life at the turn of the 21st century*. New York: St. Martin's Press.

CNN. 1996. First electronic computer turns 50: Gore re-boots ENIAC. February 14, 1996. Web posted at: 2:50 p.m. EST. From Correspondent Al Hinman. URL http://www.cnn.com/TECH/9602/eniac/ [viewed November 15, 2004]

E-LIS home page. 2005. URL: http://eprints.rclis.org/ [viewed July 1, 2005]

Ex Libris home page. n.d. URL: http://www.exlibrisgroup.com/digitool.htm [viewed November 5, 2005]

Fez: Project Overview. URL: http://espace.library.uq.edu.au/documentation/ [viewed November 5, 2005]

Gladney, Henry M., Fred Mintzer, and Fabio Schiattarella. 1997. Safeguarding Digital Library Contents and Users Digital Images of Treasured Antiquities. *D-Lib Magazine*: July/August. URL: http://www.dlib.org/dlib/july97/vatican/07gladney.html [viewed July 5, 2005]

Harnad, Stevan and Tim Brody. 2004. Comparing the Impact of Open Access (OA) vs. Non-OA Articles in the Same Journals. *D-Lib Magazine* 10(6). URL: http://www.dlib.org/dlib/june04/harnad/06harnad.html [viewed July 5, 2005]

Hopkins faults safety lapses: Panel says volunteer likely died from drug used in asthma study; Board, researcher blamed by Jonathan Bor and Tom Pelton, *Baltimore Sun*, Originally published July 17, 2001. URL: http://www.baltimoresun.com/balte.md.hopkins 17jul17.story [accessed November 15, 2004]

Huwe, Terence K. 2005. My Three Wishes for Digital Repositories. *Computers in Libraries*: 25(4): 32–34.

Koehler, W. 2004. A longitudinal study of Web pages continued: a report after six years. *Information Research* 9(2): paper 174. URL: http://InformationR.net/ir/92/paper174.html [viewed July 2, 2005].

Lynch, Clifford A. 2003. Institutional Repositories: Essential Infrastructure for Scholarship in the Digital Age *ARL Newsletter* 226 (February): 1–7. URL: http://www.arl.org/newsltr/226/ir.html [viewed July 5, 2005]

Marcum, Deanna B. and Kenney, Anne R. 2002. The Preservation of our Brittle Books Must Also Preserve Access. *Chronicle of Higher Education* 48(26, March 8)

National Archives (U.K.) 2004. *Guidelines on developing a policy for managing email*. URL: http://www.nationalarchives.gov.uk/electronicrecords/advice/pdf/ managing_emails.pdf. [Viewed July 2, 2005].

National Library of Australia. n.d. *Preserving access to digital information*. URL: http://www.nla.gov.au/padi/index.html [viewed June 10, 2005]

Rothenberg, Jeff. 1995. Ensuring the Longevity of Digital Documents. *Scientific American* 272(1): 42–47.

Smith, MacKenzie et al. 2003. DSpace: An open source dynamic digital repository. *D-Lib Magazine* 9(1). URL: http://www.dlib.org/dlib/january03/smith/01smith.html [viewed July 2, 2005]

VTLS. VITAL home page. URL: http://www.vtls.com/Products/vital.shtml [viewed November 2, 2005]

Yan Han. 2004. Digital content management: the search for a content management system. *Library Hi Tech* 22(4): 355 – 365.

Editorial history:
paper received 7 July 2005;
final version received 27 November 2005;
accepted 28 November 2005.