Lasting Impact:
Sustainability of Disciplinary Repositories

Ricky Erway

Senior Program Officer
OCLC Research

Lasting Impact: Sustainability of Disciplinary Repositories
Ricky Erway, for OCLC Research

Suggested citation:
Erway, Ricky. 2012. Lasting Impact: Sustainability of Disciplinary Repositories. Dublin, Ohio:
OCLC Research. http://www.oclc.org/research/publications/library/2012/2012-03.pdf.

Contents

Tables

Introduction

Librarians need to be familiar with the evolving aspects of scholarly communication and the changing scholarly record. One component of that is the role of repositories. It's crucial for anyone working in a research library to understand the repository landscape, both to advise researchers on where to look for information and how to disseminate their own research articles. Librarians should appreciate the nature of the leading disciplinary repositories and have a sense of their motivations, their scope, and how they operate. Before getting involved with a disciplinary repository, they should be familiar with the risks and opportunities in depending on the repository and, most importantly, they need to know if the repository has a sustainable model. For a library considering starting a disciplinary repository or taking on the operation of an existing one, these considerations are essential.

Definitions

For this report, I define *disciplinary repositories* as places where the findings of research in a particular field of study are made accessible. Any researcher in the field must be eligible for inclusion (i.e., the repository is not limited based on institution, nationality, journal publisher, or funding body). Disciplinary repositories provide access to journal articles (pre- or post-print), but may also provide access to other types of information and offer other services. Some of these repositories provide searching of metadata only and then offer links to the full text stored elsewhere; others offer searching of metadata and full text—and in those cases, the full text may be on the repository site or may be linked to elsewhere.

Disciplinary repositories are not uniform in nature. Some disciplinary repositories accept additional types of content, such as datasets, presentations, and working papers. Some offer additional services, like RSS feeds, collaboration tools, and bibliography and curriculum vitae services. At a minimum, a disciplinary repository provides access to research articles in a particular field. As suggested above, some repositories store the articles and some merely link to them in other locations. Most researchers don't know or care where articles are stored and therefore don't differentiate between sties that consist only of metadata that points to articles elsewhere and true repositories that actually contain articles. For this reason, I included both of these types of repositories in this review.

I define *sustainability* as the ability to keep an already successful repository running into the future. Sustainability in this case does not take into account past costs or return on investment.

The Patchwork Landscape of Research Repositories

Nearly every university has, or feels the need to have, an institutional repository. The usual purposes of an institutional repository are to preserve and to make accessible the articles written by those at the university. An ancillary use is to use the contents of the institutional repository to quantify the research output of the university. Some institutional repositories contain a variety of other materials beyond research articles, such as gray literature, institutional records, and digital library collections. Universities have met with varying degrees of success in realizing the goal of having all research outputs represented in their institutional repositories. Where there is a mandate to deposit, especially in nations where institutional funding is allocated on the basis of the research output assessment, there is greater compliance. For the most part, however, researchers aren't very interested in the institutional repository.

There are disciplinary repositories in many fields. They co-locate research articles in a particular subject area and are often researcher-initiated. These may be run by a scholarly society, a governmental agency, a university library, a commercial entity, or by researchers themselves. The usual purposes of a disciplinary repository are to provide a place for a researcher to find other researchers' work in a given field and to provide a way for a researcher to improve the chances that his work will be found by others in the field. Disciplinary repositories tend to be more successful than institutional repositories in attracting researchers to use and contribute to the repository.

Other components of the landscape are the repositories run by journal publishers, funding agencies, and those run by and for a particular nation. These tend to be populated by mandate, and have varying degrees of use.

Researchers are just as passionate about disciplinary repositories as institutions are about institutional repositories. The central repository for a researcher's field of study is where he goes for information, to see what's been published, and to look for collaborators. It's only natural that he would think of the same location when it comes time for him to deposit his work.

The landscape of repositories has much overlap and many gaps. Collaborative work by researchers at multiple institutions and from multiple nations may appear in many

repositories. A scientific journal article reporting the outcomes of grant-funded research may be deposited as a preprint in one or more disciplinary repositories, in the journal repository when it's published, in the funder's repository and in the institutional repository as a post-print. In some fields, like high-energy physics, researchers have been sharing preprints among themselves for decades, long before most librarians thought about digital repositories at all. Other fields aren't well-represented by repositories and some fields may primarily communicate through conference proceedings or another form of discourse. For researchers in disciplines that do not have disciplinary repositories, the institutional repository provides some recourse.

A clean model would have researchers depositing their work in their institutional repository where it would be preserved and made accessible to users and to harvesters. In this way, all articles would be preserved and accessible—and disciplinary repositories could harvest the appropriate articles and make them more easily accessible to the researchers in their field. The institutional repositories could also feed other services, such as expertise profiling systems, research assessment, or creation of faculty web pages.

Such a clean model may not be plausible because, even with mandates and with library staff willing to facilitate the deposit, institutional repositories are underutilized—and not all institutions have developed the capability to ensure the long-term preservation of the repository content. Disciplinary repositories have a magnetic effect on researchers that institutional repositories lack. The clean model may not even be desirable, because disciplinary repositories can develop discipline-specific functionality whereas institutional repositories would not be able to do that for all the disciplines represented. As there are large numbers of disciplinary repositories growing in size and in importance to researchers and as it is clear that disciplinary repositories will remain an important part of the landscape at least for the near future, librarians offering research support services should be well-versed about the repository landscape.

Financial support for disciplinary repositories can come from a variety of sources. Grant funding often covers the start-up costs, institutions often provide pro bono services, and volunteers often serve as editors. Universities are sometimes motivated to host and manage a disciplinary repository to reflect and build upon a center of excellence. In some cases libraries are involved in the development and sometimes libraries support the ongoing operations of a disciplinary repository, but for the most part, disciplinary repositories exist outside the library. It is important, however, for librarians to understand something about how these disciplinary repositories have come to be, how they fit into the landscape, and how they persist.

Methodology

I began this very casual treatment by reviewing the Consejo Superior de Investigaciones Científicas (CSIC) Webometrics July 2011 ranking of repositories (CCHS-CSIC 2011). Of the top one hundred repositories, I deemed 86 to be institutional repositories or otherwise limited to a particular population. The 14 I identified as disciplinary repositories were included in my initial investigation. These were supplemented by others that were named as a "top ten repository" by Adamick and Reznik-Zellen (A&R-Z) (2010), and a few additional repositories of interest rounded out the set. After a review of each of the 23 repositories, I selected seven to profile: AgEcon Search, arXiv, Economists Online, E-LIS, RePEc, PubMed Central, and SSRN.

Table 1. Repositories considered in this investigation

CSIC	A&R-Z	Profiled	Disciplinary Repository
5			ADS—Astrophysics Data System
59	6	x	AgEcon Search—Research in Agricultural & Applied Economics
	8		Archive of European Integration
2	4	x	arXiv.org e-Print Archive
3	2		CiteseerX
66			Cogprints Cognitive Sciences EPrint Archive
74			Cryptology ePrint Archive
71			Depot Erudit—Érudit Documents and Data Repository
		x	Economists Online
33	9	x	E-LIS: E-Prints in Library & Information Science
			EthicShare
			InterNano—Resources for Nanomanufacturing
10			Munich Personal RePEc Archive
45	10		Organic Eprints
			PASCAL2—Pattern Analysis/Statistical Modelling/Computational Learning
44			PEDOCS—Das Fachportal Pädagogik
64			Philosophy of Science Archive
			PhilPapers—Online Research in Philosophy
	7		Policy Archive
	1	x	PubMed Central
4	3	x	RePEc—Research Papers in Economics
			SSOAR—Social Science Open Access Repository
1	5	x	SSRN—Social Science Research Network

The repositories profiled serve as examples of a variety of business models and approaches to sustainability. This is not a scientific study and I did not set out to compare all aspects of disciplinary repositories. What follows is a thumbnail profile of each of the seven repositories with a brief description of their approach to sustainability. The report concludes with an overview of the sustainability models and a look at ways to cut costs and supplement revenue in order to maintain disciplinary repositories as an integral part of the landscape.

Repository Profiles

AgEcon Search: Research in Agricultural & Applied Economics

Primary business model: Institutional support

AgEcon Search[1] is hosted by the Department of Applied Economics and University Libraries at the University of Minnesota, with cooperation from the Agricultural and Applied Economics Association. It began in 1995 as a gopher service with WordPerfect documents. It contains nearly 50,000 documents covering agricultural economics and sub-disciplines such as agribusiness, food supply, natural resource economics, environmental economics, policy issues, agricultural trade, and economic development. AgEcon Search contains working papers, conference papers and journal articles (also book chapters, books, theses, dissertations, reports, and other documents) that are part of a series or conference or journal submitted by contributing organizations, such as academic departments, government agencies, non-government organizations, or professional associations. About 50 journals are included and all documents are full text and in PDF. AgEcon Search commits to providing long-term access and preservation. A free weekly update e-mail service that lists selected new papers is available and statistics are listed for each document and each contributing organization. AgEcon Search is moving from DSpace to Fedora and feeds agricultural economics content to the Agriculture Network Information Center (AgNIC).

There are no charges for using the service nor to organizations that designate one person to fill out submission forms and upload papers. There are small charges per paper for groups that have each author submit their own papers, or that have University of Minnesota staff do the inputting. The Department of Applied Economics and the University Libraries at the University of Minnesota provide hosting services and two staff members. It is co-sponsored by the Agricultural and Applied Economics Association. If AgEcon Search is closed down, the database will be transferred to another appropriate archive.

arXiv.org

Primary business model: Use-based institutional contributions

arXiv[2] is hosted by Cornell University Library. It began in 1991 at Los Alamos National Laboratory, where it was hosted for the first ten years. It contains 735,000 full-text articles, covering theoretical high-energy physics and other active research fields of physics, mathematics, nonlinear sciences, computer science, statistics, and the physics-related parts of biology and finance. Authors deposit approximately 75,000 articles a year. Some of the articles have small associated objects or link to datasets stored in the Data Conservancy or elsewhere. Moderators conduct minimal filtering of incoming preprints to maintain basic quality control for appropriateness to the designated subject areas. There are roughly one million full-text downloads by about 400,000 distinct users every week. arXiv is retooling to layer arXiv-specific capability over generic, open-source repository software, Invenio, which will allow better interoperability with ADS, the Astrophysics Data System, and INSPIRE, a repository combining three main high-energy physics resources.

There are no fees to download or deposit articles. arXiv makes use of volunteer external moderators. Mirror sites are supported by their hosts. Until recently, arXiv has been supported by its host institutions, with some NSF funding. arXiv's budget for 2012 is $589,000, of which over 80% is salaries. In order to maintain the service, Cornell seeks recurring annual contributions from the libraries of the 200 institutions that request the most downloads. They succeeded in making their target in 2010 and exceeded their target in 2011, allowing Cornell's support for arXiv to level off at 15%. The three-year, short-term contribution model has been successful and, with Cornell's contributions, has resulted in sufficient funds to cover all direct expenses for 2010 and 2011 as well as creating a contingency fund. They expect to announce a funding model for 2013-2017 by June 2012, as well as plans to transition to a collaboratively governed, community-supported resource. The long-term plan will likely incorporate Cornell support, institutional subsidies, and one or more of: scholarly society contribution, endowment, or funding from NSF or other funding agencies. arXiv may offer some additional services to supporting institutions that do not compromise commitments to open access and free submission.

EconomistsOnline

Primary business model: Support via consortium dues

EconomistsOnline[3] is hosted by Tilburg University. EO gets its information (harvests metadata and crawls digital files) from institutional repositories and makes them full-text searchable in a single location. EO includes academic publications and datasets, working papers, conference proceedings, reports, government publications, statistics, and theses. EO is predominantly metadata, many with links to open access full text, and an index of the full text. If documents are image only, EO does OCR in order to index the text. There are over 1.1M bibliographic references, 40,000 full-text documents, and 100 datasets. It offers facetted search, multilingual searching, and RSS feeds. It bases its selection on the output of economists at participating institutions (31 institutions from 21 countries) and currently contains profiles and complete publication lists for over 1,000 scholars. EO works with RePEc (in fact 1 million of their records come from RePEc), but was unable to work out a deal with SSRN, due to their business model. In some cases, allowing EO to harvest and subsequently place the content in a RePEc archive precludes the need for some institutions to maintain a separate RePEc archive. EO uses the MERESCO software suite and Dataverse for datasets.

EconomistsOnline is run by the Nereus consortium and was co-funded by the European Union, while it was in its project phase. The Nereus consortium provides the financial means to sustain the service. Post-project sustainability is guaranteed by the commitment of partner institutions to develop and sustain their respective institutional repositories. All contributing institutions have agreed to assign 0.25 FTE of the information specialist staff to work on acquiring new content. Tilburg University runs the gateway as part of its strategic plan for international collaboration and support. EO is developing the service with more partners, extended services, an expanded network, and a more global approach. All Nereus partners pay a yearly fee which is used for the maintenance of the EconomistsOnline service.

E-LIS: E-Prints in Library & Information Science

**Primary business model: Distributed network of volunteers**

E-LIS[4] covers the fields of library, information science, and technology and began in 2003. It includes journal and conference contributions, books and book chapters, theses, and research reports and working papers in full text, whether published or not. It has 12715 documents and gets 140 per month (in 2009). Authors self-deposit though E-LIS can mediate deposit. Editors in 40 countries control metadata quality. E-LIS also offers multilingual support, statistical info (usage per document), downloads into EndNote, and other services. E-LIS migrated from EPrints to DSpace.

Initial funding for E-LIS was from the Spanish Ministry of Culture. There are no fees to download or deposit articles. E-LIS is managed by CIEPI (International Centre for Research in Information Strategy and Development) and depends on volunteers in three organizational structures: Administration handles policy and organizational development; Editors edit and promote (up to three in a country); and Technology handles software implementation and OAI. They get support from the Agricultural Information Management Standards (AIMS) team, which is facilitated by United Nations bodies. Hosting and technical maintenance are pro bono and currently performed by Academic E-Publishing Infrastructures - Consorzio Interuniversitario Lombardo per L'Elaborazione Automatica (AePIC CILEA, a non-profit Italian university supercomputing consortium).

PubMed Central

Primary business model: Federal government funded
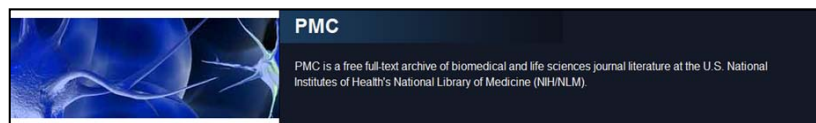
PubMed Central[5] (PMC) is hosted by the U.S. National Institutes of Health's National Library of Medicine (NIH/NLM) through its National Center for Biotechnology Information (NCBI). It was launched in February 2000 and covers biomedical and life sciences, biology, bio-chemistry, and health and medicine. It accepts deposits of full journal issues from about 1000 journals, plus selected articles from thousands of other journals. There are currently 2.37 million full-text articles, growing about 10% annually. Content is deposited by participating publishers (and authors of manuscripts that are covered by the public access policies of certain approved

funding agencies). PMC accepts material from any life sciences journal that meets NLM's scientific and technical standards. Publishers can delay access for up to a year or more. PubMed Central provides preservation services and has OAI and FTP services that allow access to all the metadata and to full text for a subset of the articles with appropriate licenses. PMC provides publishers with export of all their data for inclusion elsewhere and provides usage stats. PMC is based on software developed by NCBI. Most of the PMC articles have a corresponding entry in PubMed, the database of citations and abstracts for millions of articles from thousands of journals, which includes links to full-text articles at several thousand journal websites as well as to most of the articles in PubMed Central.

There are no fees to download or deposit articles. NIH provides funding and the service is managed by NCBI. It has counterparts in the UK (UK PubMed Central) and in Canada (PubMed Central Canada).

RePEc: Research Papers in Economics

Primary business model: Decentralized arrangement

Various components of RePEc[6] are hosted by the Federal Reserve Bank of St. Louis, Swedish Business School Orebro University, Munich University Library, SUNY Oswego, Valencian Economic Research Institute (Spain), and Nereus—and Boston College hosts the domain. It was created as Gopher services in 1993 and became RePEc in 1997. RePEc contains working papers, journal articles, software components, books and chapter listings in the areas of economics and business, as well as author contact and publication listings and institutional contact listings. RePEc is an open-source bibliographic data collection and does not manage documents; to be included; documents must be stored in up-to-date institutional repositories (IRs) in RePEc format. RePEc harvests from 1397 archives, currently containing metadata for 1.1M items, the vast majority of which link to documents available on-line. They estimate about 700K downloads per month. RePEc also offers citation analysis, logging service, author service, download counts, listing of top articles, as well as alerts, a mail list, and a blog. They are attempting to offer an alternative form of certification to that provided by traditional journals. The American Economic Association (AEA) selected RePEc to serve as its exclusive source providing bibliographic and abstract data for working papers indexed in EconLit. In return, the AEA will encourage all major research institutions to maintain an up-to-date repository of working papers so that RePEc has the most up-to-date working papers. RePEc runs on local software and harvests from institutions via its own protocol, but offers its content via OAI.

There are no fees to include metadata or search metadata (links to articles elsewhere may be free or fee-based). In its earliest days, RePEc received support from the Joint Information Systems Committee (JISC) of the UK Higher Education Funding Councils, as part of its Electronic Libraries Programme (eLib). RePEc is based on volunteerism and community support; all its content is decentralized and hosted at participating departments and institutions. There is no central management and, while some individuals have taken on responsibilities for certain areas of the site, there is no formal structure.

SSRN: Social Science Research Network

Primary business model: Commercial "freemium" service

SSRN[7] is hosted by Social Science Electronic Publishing, Inc. an independent corporation based in Rochester NY. It began in 1994 and focuses on social science and humanities research. The SSRN eLibrary database includes over 1000 eJournal classifications in eighteen specialized research networks in the social sciences and five in the humanities. The eLibrary includes 385,000 abstracts and associated metadata from 180,000 authors. Over 316,000 of the abstracts have full-text papers available for download in PDF format. 60,000 new papers were submitted last year. 1800 journals are represented as Partners in Publishing, which means they have submitted working papers or the associated journal has given permission for inclusion of its abstracts in the eLibrary. SSRN has over 1.4 million users who downloaded 8.5 million full-text PDFs last year. Every submitted paper is reviewed by SSRN staff to ensure it is part of the scholarly discourse and appropriately classified in up to 12 eJournals across multiple networks. SSRN provides author contact information, extracts references and footnotes from the PDF, and provides article level metrics (downloads, citations, and Eigenfactor™) to facilitate a broader understanding of scholarly impact and enhance user searching. SSRN is built on customized software and has mirrored paper repositories at ECGI in Belgium, University of Chicago Booth School of Business, Korea University, and Stanford Law School.

It is free to submit research, access abstracts and associated metadata, and download the full text PDF (unless the publisher submitted the paper and charges for access). SSRN is a closely-held, for-profit company, which has self-funded the service. It generates revenue from individual and organizational site subscriptions, which allow subscribers to receive regular e-mails with the newest abstracts and metadata, and Partner in Publishing services, which provide scholarly research management and distribution services for schools, departments, research organizations, and conferences. Within Partners in Publishing, SSRN provides

discipline focused Research Paper Series for schools and departments; effectively, a back-office SSRN designed to increase exposure, simplify access, and expand readership for an organization's research. They have Google ads on some abstract pages and offer purchase of hard bound copies of PDFs. SSRN also offers a Professional Jobs and Announcements service that is free to subscribers. The fees for this service are paid by the university, company, or conference organizer.

Table 2. Overview of the profiled repositories

Disciplinary Repository	Country	Launch Date	Size	Metadata / Documents	Acquisition Approach	Primary Business Model
AgEcon Search – Research in Agricultural & Applied Economics	US	1995	Almost 40K	Documents	Publishers and organizations deposit	Institutional support
arXiv.org e-Print Archive	US	1991	735K	Documents	Authors deposit	Use-based institutional contributions
Economists Online	Netherlands	2010	1.1M/40K	Metadata / documents	Harvests metadata from IRs; indexes full text	Support via consortium dues
E-LIS: E-Prints in Library & Information Science	Italy	2003	12.7K	Documents	Authors deposit	Distributed network of volunteers
PubMed Central	US	2000	2.4M	Documents	Publishers deposit + some authors deposit	Federal government funding
RePEc –Research Papers in Economics	Netherlands, US, Sweden, Germany, Spain	1997	1.135M	Metadata	Harvests metadata from IRs	Decentralized arrangement
SSRN –Social Science Research Network	US	1994	385K/316K	Metadata / documents	Authors and publishers deposit	Commercial "freemium" service

Sustainability

Even if a library is not thinking about running a disciplinary repository, it will likely end up depending on disciplinary repositories to help support its university's research mission, so it's important for librarians to understand repositories' business models to be able to judge their long-term reliability.

While many think of disciplinary repositories as being free, there are always real costs and somebody pays them. In 2010, Alma Swan gave four examples of running costs of institutional repositories that ranged from €30,000–€225,000 per year. Perhaps the

longest-running, most efficient, and most heavily-used repository, arXiv, has costs of nearly half a million dollars a year and calculates the cost per download to be 1.4 cents or the cost per deposit to be just under $7. For most other repositories, the unit costs are likely to be significantly higher.

There are many ways to cover the costs associated with a repository and many ways to manage those costs. Some possible models that I did not see in my sample are: personal or institutional access subscription (often used by journals); contribution by author payment (often used by "gold road" open access publishers); and support through endowment (used by the lucky few).

The dominant funding models (some had secondary sources of funding) described in the profiles include:

- Institutional support

- Use-based institutional contributions

- Support via consortium dues

- Distributed network of volunteers

- Federal government funding

- Decentralized arrangement

- Commercial "freemium" service (basic access is free; value-added services for fee)

Based on the varying approaches to sustainability taken by disciplinary repositories, it is clear there is no single answer and, in fact, most repositories use a combination of approaches. The most straight-forward approach is to have 100% support from government funding or from an endowment, but you've either got it or you don't—and most don't. Having an institution that is committed to funding, hosting, and growing the service is a close second—second only because the institution-supported repository might be more likely to fall victim to budget cuts. But there are compelling combinations, such as having a grant to cover startup costs, an institution willing to shoulder the costs of hosting the service, and a team of dedicated volunteers.

There is also a variety of approaches to managing costs. Just hosting metadata instead of also storing and serving the digital documents can keep costs down. Harvesting metadata from other sources can be more efficient than getting individual deposits. Getting deposits from journals can be easier than dealing with individual authors. Not committing

to store and preserve documents avoids a substantial undertaking. Using open source software that is supported and maintained elsewhere can be cheaper than building and maintaining your own platform.

Repositories do scale—that is, the unit costs decrease as contributions and use increase. Scaling is possible when the repository is successful and has high usage. There are a number of factors that contribute to a repository's success, such as:

- Prominent people in the field promote it

- Partnerships with publishers and societies

- Quality control

- Visibility in web search results

- Added services, like researcher bibliographies or e-mail alerts

Base funding and cost-containment methods can be supplemented by measures to bring in additional revenue:

- Freemium user services

- Custom services and consulting for other repositories

- Corporate or association sponsorship

- Grant funding for upgrades or for development of new features

- Advertising

Of course, with any model or combination of approaches, a service continuity plan is necessary. What will happen to the content if the funding (or volunteerism) should cease? Who would keep it alive? Could it be turned over to another entity? Contributors and investors should want to know this up front.

While many debate the relative merits of institutional repositories and disciplinary repositories, I think they are both here to stay. Librarians should consider the roles and possible relationships between the two types of repositories and how they might work together to achieve mutual goals. Librarians should ask themselves, while many disciplinary repositories harvest from institutional repositories, could the institutional repository harvest from disciplinary repositories? What functions does each best support? How can library staff best advise their researchers? Should libraries get involved in planning and running

disciplinary repositories? And most importantly, how can universities, libraries, and the researchers they serve make the most of the entire patchwork repository landscape?

Acknowledgements

Notes

1. AgEcon Search: http://ageconsearch.umn.edu

2. arXiv: http://arxiv.org

3. EconomistsOnline: http://www.economistsonline.org

4. E-LIS: http://eprints.rclis.org

5. PubMed Central: http://www.ncbi.nlm.nih.gov/pmc

6. RePEc: http://repec.org

7. SSRN: http://ssrn.com

References

Adamick, Jessica and Rebecca Reznik-Zellen. 2010. "Representation and Recognition of Subject Repositories." *D-Lib Magazine* 16 (9/10). http://www.dlib.org/dlib/september10/adamick/09adamick.html.

CCHS-CSIC (Centro de Ciencias Humanas y Sociales—Consejo Superior de Investigaciones Cientificas). 2011. "Top Repositories." *Ranking Web of World Repositories*. Accessed September 2011. http://repositories.webometrics.info/toprep.asp.

Swan, Alma. 2010. "Business Planning for Digital Repositories." In *Business Planning for digital Libraries: International Approaches*, edited by Mel Collier. Leuven, Belgium: Leuven University Press. http://eprints.ecs.soton.ac.uk/21470/.