https://doi.org/10.26069/greynet-2026-000.613-tgj
ISSN 1574-1796

An International Journal on Grey Literature

Summer 2005 - Volume 1, Number 2

'REPOSITORIES – HOME2GREY'

www.textrelease.com GreyNet Grey Literature Network Service www.greynet.org

https://doi.org/10.26069/greynet-2026-000.613-tgj

The Grey Journal
An International Journal on Grey Literature

COLOPHON

CIP
The Grey Journal : An international journal on grey literature / D.J. Farace (Journal Editor); J. Frantzen (Technical Editor) ; GreyNet, Grey Literature Network Service. - Amsterdam: TextRelease, Volume 1, Number 1, Spring 2005.
This serial publication appears three times a year - in spring, summer, and autumn. Each issue in a volume is thematic and deals with one or more related topics in the field of grey literature. The Grey Journal appears both in print and electronic formats.
ISSN 1574-1796 (Print)
ISSN 1574-180X (E-Print/PDF)
ISSN 1574-9320 (CD-Rom)

Subscription Information
Subscription Rates: €85 individual,
€200 institutional; rates include postage & handling.
Reprint Service, Back Issues, Advertising and Inserts contact:

TextRelease
Beysterveld 251
1083 KE Amsterdam
Netherlands
T/F +31 (0) 20 672.1217
info@textrelease.com
http://www.textrelease.com

About TGJ
The Grey Journal is a flagship journal for the grey literature community. It crosses continents, disciplines, and sectors both public and private. The Grey Journal not only deals with the topic of grey literature but also is itself a document type that is classified as grey literature. It is akin to other grey serial publications, such as conference proceedings, reports, working papers, etcetera.
The Grey Journal is geared to Colleges and Schools of Library and Information Studies, as well as, information professionals, who produce, publish, process, manage, disseminate, and use grey literature e.g. researchers, editors, librarians, documentalists, archivists, journalists, intermediaries, etc.
Grey Literature is defined as "information produced on all levels of government, academics, business and industry in electronic and print formats not controlled by commercial publishing *i.e. where publishing is not the primary activity of the producing body.*" (Luxembourg 1997; expanded in New York, 2004)

About *GreyNet*
The Grey Literature Network Service was founded in 1993. The goal of GreyNet is to facilitate dialog, research, and communication between persons and organizations in the field of grey literature. GreyNet further seeks to identify and distribute information on and about grey literature in networked environments. Its main activities include the International Conference Series on Grey Literature, the creation and maintenance of web-based resources, a moderated Listserv, The Grey Journal, etc.

https://doi.org/10.26069/greynet-2026-000.613-tgj

Contents
"REPOSITORIES – HOME2GREY"

https://doi.org/10.26069/greynet-2026-000.613-tgj

"REPOSITORIES – HOME2GREY"

Last year, in an online survey leading up to the Sixth International Conference on Grey Literature, one of the survey items questioned if grey literature collections are better managed by institutional (centralized) rather than disciplinary (decentralized) repositories? Roughly a third of the respondents agreed, a third disagreed, and a third were uncertain. During the actual conference in New York, a plenary session dealt with the topic of institutional repositories (IR) and much of that content is found in this journal issue.

In the six features articles, repositories managed by both government and academic institutions are openly and candidly discussed. The upside and downside, the success and pitfalls, the aims and future of grey literature repositories are found in these pages. In the first article, Gelfand *(UCI, United States)* is extremely upbeat in her portrayal of IR "Capturing this content institutionally adds prestige and visibility to resources that without this institutional affiliation may not have peer review, be available digitally and thus remotely, and have perpetual access." In the next three articles, case studies are used in dealing with the topic. LaFleur and Rupp *(Cornell University, United States)* designed a pilot project in which access to Cornell-produced conference proceedings could be made available online using the D-Space Institutional Repository system. They loaded the set of proceedings they had identified into D-Space and created metadata to facilitate searching for the proceedings within the system. Claerebout *(Université des Sciences et Technologies de Lille, France)* discusses Grisemine, a digital library destined to be a mine of grey literature. Since its creation in 2001, this institutional repository has gradually grown rich through the selection of teaching and research documents including theses, dissertations, course materials, conference papers, scientific reports, etc. Japzon and Anderson *(NASA Goddard Flight Center; Information International Associates, IIa)* explain Goddard Library's development of DAS, Digital Archiving System, which is a prototype infrastructure for creating a combined metadata repository that allows metadata for heterogeneous digital objects to be searched with a single search mechanism. Their example used is the Wallop's Balloon Technology documents, which they hold could serve as a model for other NASA and/or government projects. Dijk *(Netherlands Institute for Scientific Information Services, NIWI/KNAW)* moves from a document level to a systems level approach. As part of the Dutch DARE (Digital Academic Repositories) Programme, the OA-x project has been set up to enable researchers and administrators of digital archives to be able to unlock, edit, supplement, combine and archive metadata and data (objects) in digital repositories. A protocol for harvesting and uploading objects has been developed in this project. And, in the article by Barrueco and Krichel *(University of Valencia, Spain; Palmer School of Library and Information Science, United States)* RePEc (Research Papers in Economics), the largest decentralized non-commercial academic digital library in the world is used to test CitEc, an autonomous citation index. Both its architecture and performance are analyzed in order to determine if the system has the quality required for information retrieval and for the extraction of bibliometric indicators.

I take this opportunity to welcome Prof. Gretta E. Siegel *(Portland State University)* as an Associate Editor for The Grey Journal. And, I would like to call to the readers' attention that the next issue of TGJ deals with grey literature and education.

Dominic J. Farace, Journal Editor
journal@greynet.org

https://doi.org/10.26069/greynet-2026-000.613-tgj

"Knock, Knock:" Are Institutional Repositories a Home for Grey Literature?*

Julia Gelfand (United States)

Abstract

Academic and special libraries are eagerly as well as reluctantly joining the bandwagon to participate in institutional repositories. The young and growing collection of the University of California Institutional Repository hosted by the California Digital Library (CDL) contains nearly 5300 documents (see http://repositories.cdlib.org/escholarship/). This paper will analyze the contents of that collection in terms of levels of greyness. Content comes from 9 different campuses composing the University of California's Research Units, Centers, and Departments and includes working papers, research content, journals and peer-reviewed series. This author has developed a five-point scale that identifies and describes the range of content to conclude the extent that this example of an institutional repository is grey. Institutional Repositories have different collection and review policies and this will be noted. Capturing this content institutionally adds prestige and visibility to resources that without this institutional affiliation may not have peer review, be available digitally and thus remotely, and have perpetual access. A conclusion will be made whether this model of institutional repository supports a new publishing method for renewed life in grey literature.

Introduction

I will begin the discussion about how institutional repositories have taken on a new role in higher education and scholarly publishing and what implications this has had and may have for grey literature by using the experience of the University of California eScholarship program. An international conference was just sponsored in November 2004 by the Association of Research Libraries (ARL) and the Scholarly Publishing Academic Resources Coalition (SPARC & SPARC Europe) on the next wave of institutional repositories (IR). Meetings of library leaders around the world have taken place over the past few years to launch the emergence of institutional repositories and it is the opinion of this author that it is a way of giving legitimacy to grey literature, even if only scant reference to that has been articulated in the literature or in public forums until now. So knock, knock, someone is really home now. Clifford Lynch writes about this in early 2003 by stating,

"The development of institutional repositories emerged as a new strategy that allows universities to apply serious, systematic leverage to accelerate changes taking place in scholarship and scholarly communication, both moving beyond their historic relatively passive role of supporting established publishers in modernizing scholarly publishing through the licensing of digital content and also scaling up beyond ad-hoc alliances, partnerships and support arrangements with a few select faculty pioneers exploring more transformative new uses of the digital medium."

This powerful and eloquent statement does all to confirm the definition of grey literature by inference and establish a new home for it. Today, still in its infancy, the institutional repository by all accounts is more than a nursery; it is a palace with vast Real-estate as it expands its horizons and hospitality to even more forms of information products and reaffirms its essence and stature in the academic community. Grey literature, as tweaked and redefined at the GL 1997 Conference in Luxembourg is defined as "that which is produced on all levels of government, academics, business, and industry in print and electronic formats, but which is not controlled by commercial publishers."[1] For clarity and contrast, the Institutional Repository is best defined by Lynch as "a set of services that a university offers to the members of its community for the management and dissemination of digital materials created by the institutions and its community members."[2]

Grey literature differs from commercial publications in that it is not based solely or even principally on an economic model, but rather on a communication model, which we also now describe as scholarly communications. This confirms the issues

* First published in the GL6 Conference Proceedings, January 2005

many others and I have written about concerning why grey literature is difficult in collection development and collection management realms. Until electronic publishing and the web was stable it was difficult to identify, obtain and bibliographically describe.[3] Still, we could benefit from more standards to support even better access.

Two years ago, Roy Tennant of the California Digital Library (CDL) wrote a very concise and informative article on the nuts and bolts of institutional repositories, introducing and explaining the software that ranges from open source to commercial varieties that defines the IR operation. In that short piece, he covers implementation models, keys for access, economic models, subject terminologies and some posting and removal policies. He also has a short concluding paragraph entitled, "From grey to black and white," where he concludes, "They provide much better access to a literature than has ever previously been possible and should be a no-brainer for most academic institutions."[4]

This reference to grey literature and the remainder of the context of Tennant's paper leads me to conclude that new learning communities have formed or will form as the result of institutional repositories being established and maturing with the depth and scope of their content. It is my speculation that the IRs will really make a difference in higher education because the content has increased legitimacy, offers simpler and better access, has gone through peer review, promotes local institutional research and output and is in the format our contemporary users prefer.

The literature on learning communities goes back nearly two decades now and incorporates lots of principles from the early work of Ernest Boyer, Peter Senge and many others. We must ask ourselves what types of learning communities are most relevant to institutional repositories and I believe a consensus would emerge that cross-curricular, purposeful, electronic, primary and secondary would all apply.[5] The Lenning and Ebbers volume, *The Powerful Potential of Learning Communities,* also explores the benefits of learning communities to building learning environments where the faculty benefits include: "diminished isolation, a shared purpose and cooperation…among colleagues, increased curricular integration, a fresh approach to one's discipline and increased satisfaction with their student's learning."[6] These attributes are parallel or perhaps even synonymous to criteria that librarians apply to the selection principles they use to guide their decisions about what library collections will contain, what materials will be retained where, what format of a product will be chosen and other decisions that impact the future of learning communities and libraries on both a macro and micro level, and information delivery and access as it applies to a wide spectrum of understanding and practice.

If nearly all learning communities have "two things in common with one being shared knowledge and the other, shared knowing"[7] then Gardner's list of what a learning community does fits almost seamlessly into our exploration of institutional repositories as a rightful and purposeful place for grey literature. He suggests that learning communities:

1. Incorporate and value diversity
2. Share a culture
3. Foster internal communication
4. Promote caring, trust and teamwork
5. Involve maintenance processes and governance structures that encourage participation and sharing g of leadership tasks
6. Forster the development of young people and
7. Have links with the outside world.[8]

Returning to the main thrust of my paper of how grey the California Digital Library's eScholarship collection is requires a few steps. My observation and participation in the program is limited to three vantage points: as an affiliated librarian at one of the 10 University of California campuses where I firstly direct users to resources and secondly direct faculty to publishing options, and thirdly, I am an independent user. In the first and third roles, I am searching for information that may not be in standard or traditional resources, meaning published books, journal collections, conference proceedings or such packages, where I use finding aids such as indexing and abstracting services or databases. The second role is a rather new departure for librarians. In recent years we have become increasingly comfortable with different facets of scholarly communication and are familiar with alternative options within the accepted publishing spectrum and now with new products and electronic mediums. We are also committed to promoting how economically and spatially unsustainable earlier models of scholarly publishing have become. In addition we have a commitment to digital preservation, open access, authors' rights,

cost containment and many other creative and intellectually motivated goals for scholarly publishing. UC Librarians encourage and promote the eScholarship activities and repository as a place to capture and retain the research and scholarship that is conducted by our leading faculty.

The new modes of scholarly publication within the University of California (UC) eScholarship program include: **institutional repositories** that promote pre-publication of materials and contain peer-reviewed content; **web-based publications** of digitally reformatted content and electronic editions of academic monographs of interest to both scholarly and general-interest readers. In addition, there are numerous partnerships with librarians, scholars, publishers, a wider information industry community where a special eScholarship repository has been created to support a fuller range of scholarly output from pre-publication materials to journals and peer-reviewed series, by offering the University of California departments and units direct control of publishing.[9]

You may ask and ponder why create eScholarship and an institutional repository for one university. Please remember that the University of California has 10 campuses participating and the Repository reflects a full spectrum of publishing activity from reports, peer-reviewed content, edited volumes, journals, pre-prints and in January 2005 will include post-prints. Membership is by the academic research unit and/or department, which serve as "gatekeepers" of the content and where editorial and administrative functions are distributed.[10] However interrelated eScholarship and institutional repositories are, I am just going to concentrate on the IR component for our discussion today.

The repository is searchable by:

- Campus
- Research unit, center or department
- Journals and peer-reviewed series
- Seminar series

Usage counts as of December 1, 2004 indicate that the Repository had 837,339 downloads to date and the most recent week experienced 20,235 downloads. John Ober in his presentation at the ARL/SPARC Institutional Repositories Conference documents patterns of downloads and participation from academic units.[11]

As the repository grows to currently reflect more than 5300 items I became curious about what the common and divergent elements are and this is what led me to dissect the content to determine the degree of grey literature that it contains, especially since the majority of use of the IR is from outside the UC system. Placing the cart before the horse let me share with you that this was a reasonably easy analysis because all the pieces fell into place without too much scattering. The method was a simple, quasi social science effort of describing content based on a palette of colors that represented certain matches. Content could be coded multiple times if appropriate, although only 37% of my sample had that possibility. It is important to understand the flexibility and context of the Repository Policies before any conclusions can be made because an interesting comparison for future studies is whether they are indeed sufficiently flexible. Currently, they include:

- Who can join
- Whose papers can be included in the Repository
- Appropriate submissions
- Peer-reviewed series
- Seminar series
- Removing a paper
- Author review & agreements
- Copyright[12]

Since authors retain the copyright for all content posted in the repository and the eScholarship initiative features a non-exclusive right so that the author is free to use the content in multiple way. This may be new for some faculty and guidelines such as the following passage may help inform specific practices:

If a working paper is published in a journal—either in the same form or, more commonly, in revised form—many journals allow the working paper to continue to be made available, especially when it is for educational/scholarly noncommercial use. Unfortunately, some journals do require that the working paper be removed. Others

grant exceptions for something like the eScholarship Repository; they just need to be asked. It is up to the faculty member to check the terms of their agreement with the journal to see what is allowed. Individual journal policies vary widely. The RoMEO Project (Rights MEtadata for Open archiving) has compiled a list of many journals' "Copyright Policies" about "self-archiving."

The Repository Benefits are equally important and they may be even more relevant to how we consider grey literature. Selective benefits include:

- Free to contribute for all University of California affiliates
- Promising alternative
- Increased visibility
- Usage reports
- eMail notification alerts to readers & users
- Permanence
- Global accessibility via the Open Archives Initiative (OAI)
- Ability to upload associated content
- Institutional identity
- Sophisticated searching
- High quality participants[13]

Since the content is available for all search engines to crawl, discover, and make available to their users, access is extended over the broadest range so the indexing and abstracting of content is among the least restrictive. This expands access beyond any previous distribution model that any form of grey literature has experienced in its history. Each search engine provides the output using different algorithms reflecting high use, relevancy, currency, etc. Obviously, part of this is due to the potential of electronic publishing formats and new assignments of authors' rights, but the blending of content into different types of information products creates some of the richest searching sources for relevant information. The recent release of Google Scholar (http://www.scholar.google.com) enhances access and offers different forms of competition to users and will probably include the eScholarship Repository in its output soon.

The palette of colors reflects the following spectrum and was coded to items in the repository according to the following assignments with greyness sliding from each color and blending in the middle. This rather simple methodology assigned a numerical value for each color or factor on a range of 1-5 with one being low and 5 high. This demonstrated how many examples in the repository that were reviewed had strong indicators for linking, environmental, publishing, collaboration and interdisciplinary factors.

Black – > **Linking Factors** – demonstrates relationships to author's other works (backward -> forward), to institutional colleagues, to other content and secondary sources that track citation histories, etc

Red – > **Environmental Factors** – reflects the complexity of initial acquisition, description, funding, troubleshooting and support for persistent and perpetual access tracking versions and usage, etc

Blue – > **Publishing Factors** - even though the effort is to reduce time to publication with the attributes of peer review, capturing older work that has never been shared or previously released is another goal as is the overall publishing process. In the future, citation analysis may be a part of this category.

Yellow – > **Collaborative Factors** – can be used in multiple functions, for instance in classroom teaching, scholarship, repackaging, etc

White – > **Interdisciplinary Factors** – promotes and incorporates the spirit of new and emerging work from multiple subject areas, and may include conference and seminar content, etc.

In June 2004 I reviewed a randomized sample of content in the following content areas: 50 examples of the research units/departments and 50 examples from the Seminar Series. Due to the 37% of multiple codings the total sample included 137 submissions, which is only 2% based on today's inventory or 8% in June. Sample size is very modest but the highest matches were for collaborative factors, followed by interdisciplinary factors, environmental factors, publishing factors and then linking factors.

Linking Factors	-	7%	or	9.59 items
Environmental Factors	-	18%	or	24.66 items
Publishing Factors	-	21%	or	28.77 items
Collaborative Factors	-	25%	or	34.25 items
Interdisciplinary Factors	-	29%	or	34.25 items
Total	=	100%	or	137.00 items

The range reflected more about grey literature than it did about other descriptive elements. My hypothesis that as the Repository grows at its rapid rate and there are more citations attributed to its content, and that the colors will blend even more and become less grey but more white, even with linking a problematic factor for a longer while. The reason for this is that linking depends on busy academics to provide the information and add more of their earlier work to offer a spectrum of work and to encourage other colleagues to contribute to the IR. Maturity in this category is more dependent on external factors than any of the others.

To conclude, we can say that the factors defined as linking; environmental, publishing, collaborative and interdisciplinary each describes grey literature contained in institutional repositories. There are longstanding issues including:

- transcience of grey literature
- maturation of the repository
- timely publishing
- access
- standards
- multiple formats
- and other areas that need more attention.

The players and partners will change and the content will not only multiply but absorb new technologies and discoveries to meet user expectations. The challenges that will remain for the foreseeable future include multiple formats, how to scale content, keep it easy, changing technologies, identification and access, need for a mixture of expertise to implement projects and leveraging of collective investments, communication and promotion of the content[14] But we know that Grey Literature will not be homeless again as long as institutions continue to exploit the possibilities and merits of building and refining the institutional repository concept. No longer is Grey Literature at risk, it will be in a good collegial neighborhood, and be sought after instead of being a weak commodity in a chain of information products that was previously inhospitable to it. Libraries are happier because the organizational structure of the institutional repository remains academic and scholarly and there is no fear of being lost. The palette of factors will become brighter in each category and fading grey will be a part of the past. The proliferation of institutional repositories and high usage and downloading is already testimony to this transformation. Knock, Knock, someone is certainly home.

References

[1]Dominic J. Farace, "STS Publisher/Vendor Relations Conference Discussion Group Conference Report," published in ISTL, Winter 1998. http://www.istl.org/98-winter/conference4.html

[2] Lynch, 327

[3] Julia M. Gelfand, "Academic Libraries and Collection Development: Implications for Grey Literature," Publishing Research Quarterly 13,2, Summer 1997: 15-24 and "Grey Literature Poses New Challenges for Research Libraries," Collection Management 24,1/2, 2000:137-148, and Gelfand and John L. King, "Grey Market Science: Research Libraries, Grey Literature and Legitimization of Scientific Discourse in the Internet Age," in *Socioeconomic Dimensions of Electronic Publishing Workshop Proceedings*, 1998:115-120. Paper presented at IEEE Conference, Santa Barbara, CA, April 25, 1998.

[4] Roy Tennant, "Institutional Repositories," *Library Journal*, September 15, 2002. http://www.libraryjournal.com/iondex.asp?layout=articlePrint&articleID=CA242297

[5] Oscar T. Lenning and Larry H. Ebbers, *The Powerful Potential of Learning Communities: Improving Education for the Future.* ASHE-ERIC Higher Education Report 26 #6. Washington, DC: The George Washington University Graduate School of Education and Human Development, 1999:iii.

[6] *Ibid*, iv.

[7] Vincent Tinto, "Colleges as Communities: Taking Research on Student Persistence Seriously," *Review of Higher Education* 21, 2, 1998:171 as referenced in Lenning and Ebbers, 10.

[8] John W. Gardner, "Building Community," *Kettering Review*, 1989: 73 as referenced in Lenning and Ebbers, 10.

[9] California Digital Library eScholarship website at http://www.cdlib.org/prgorams/escholarship.html

[10] Catherine Candee, "eScholarship," Presentation made at the Society for Scholarly Publishing Annual Meeting, June 4, 2004, San Francisco.

[11] John Ober, "IR Content & Service Expansion: the Case of eScholarship," Presentation at Institutional Repositories: the Next Stage Conference Nov 18-19, 2004, Washington, DC. http://www.arl.org/sparc/meetings/ir04/presentations/ober.html

[12] See http://repositories.cdlib.org/escholarship/policies.html

[13] For complete list see http://repositories.cdlib.org/scholarship/benefits.pdf

[14] Adapted from Lucia Snowhill, "Working with Government Information System-wide," Presentation made to UC Conference on Government Information, October 28, 2004.

In addition to the cited references these resources also contributed to this paper:

Carol A. Hughes, "EScholarship at the University of California: A Case Study in Sustainable Innovation for Open Access," *New Library World*, 105, 1198/1199, 2004: 118-124.

Institutional Repositories: Revealing Our Strengths. ARL/OLMS Webcast cosponsored by SPARC and CARL, June 10, 2004.

Marie E. McVeigh, "Open Access Journals in the ISI Citation Databases: Analysis of Impact Factors and Citation Patterns." White Paper from Thomson Corporation, 2004. Available at http://www.thomsonisi.com/media/presentrep/essayspdf/openaccesscitations2.pdf

The opinions expressed in this paper are solely those of the author and do not imply agreement or endorsement of the CDL staff.

https://doi.org/10.26069/greynet-2026-000.613-tgj

Making Grey Literature Available through Institutional Repositories*

Lee LaFleur and Nathan Rupp (United States)

Abstract

One of the major components of grey literature is conference proceedings. A major purpose of conferences is to disseminate research in a particular area, and this is often done through conference proceedings. However, although many libraries collect conference proceedings, they collect them mostly in print. With the advent of digital library collections—electronic journals, digital library projects, etc.—conference proceedings have been neglected. Even though the research in conference proceedings can be as vital as that in any journal article, the access to conference proceedings has not kept pace with access to other types of library resources.

In its support of the Life Sciences Group at Cornell, Mann Library identified this lack of online access to conference proceedings as a problem whose exploration and resolution might benefit faculty and researchers supported by the library. Librarians at Mann embarked on a pilot project to identify a set of Cornell conference proceedings as a candidate which should be made available online to researchers. In addition, they took advantage of Cornell's membership in the nascent D-Space Federation and used the D-Space Institutional Repository system to make the proceedings available online. They loaded the set of proceedings they had identified into D-Space and created metadata to facilitate searching for the proceedings within the system. They then demonstrated this system at the annual meeting of the conference.

The positive reception to the conference demonstration convinced the librarians involved with the project that making conference proceedings available online to researchers was a worthwhile service that should be further explored. In the process of identifying the proceedings, collecting them, and making them available online, the librarians encountered a number of issues that they would need to address as they further explored the service and determined whether or not it should be implemented as a full program.

Despite these issues in planning for a sustainable institutional repository for conference proceedings, the librarians involved felt that the feedback they had received from the researchers at the conference supported the notion of expanding the project. If the expansion of the project succeeds, it will be another means of providing access to grey literature.

Introduction

The Cornell University Library exists to support the three-fold mission of the University: research, education, and extension (Cornell is the land grand institution for the State of New York). It supports this mission through the services it provides and the material it collects; it collects material produced by the university itself as well as outside publishers. In recent years, university librarians have recognized that a significant amount of material worth collecting in support of the university's activities can be classified as "grey literature." Although grey literature has in the past been considered too ephemeral and marginal to spend time collecting, it is part of a communications process that, although separate from mainstream publishing, is just as important. Even though some information resources are created with nontraditional publishing methods does not mean that they are unimportant; often, these grey literature resources later move "into the mainstream of information products." (1) In addition to its general value in research, grey literature is particularly important in the sciences. "The quest for scientific knowledge is an evolutionary process in which every increment of new knowledge adds to, modifies, refines, or refutes earlier findings," and grey literature is an important part of this process. (2) To this end, librarians at Cornell, including librarians supporting the sciences, have determined that collecting grey literature is as important to supporting the university's activities as is collecting any other resource.

* First published in the GL6 Conference Proceedings, January 2005

A recently published Cornell University Library report surveyed the types of grey literature currently being produced within the university. These included the physics pre-prints contained in the arXiv e-print service; documents created by Cornell University Cooperative Extension; working papers produced by programs such as the Cornell Food and Nutrition Policy Program and the Center for Advanced Human Resource Studies; and proceedings produced in association with a number of conferences sponsored by Cornell.

Conference proceedings are an important type of grey literature for a number of reasons. Conferences and the proceedings that document their programs are a valuable part of the scholarly communications cycle. The proceedings from conferences document the current state of research and provide a lasting record about what took place at a particular conference. Furthermore, proceedings enable research to be shared among those who were not present at a conference and serve as a point of reference for those who were in attendance. Many research findings presented at conferences and recorded in conference proceedings are not reported anywhere else. However, conference proceedings can be difficult for researchers and information professionals to identify, locate and acquire. Many times the groups that organize conferences are small and their events are not widely publicized. Even those proceedings collected by libraries can often be difficult to access due to lack of item level indexing, frequent title changes and the irregular and informal nature of their publication.

Conference Proceedings Project

As the report surveying grey literature at Cornell was being researched, other library staff members had begun to think about how the library could provide online access to proceedings published in association with conferences held at Cornell. They designed a pilot project in which access to Cornell-produced conference proceedings could be made available online, and in doing so, identified a number of activities such a project would need to undertake. These included the identification of a set of conference proceedings with which to establish the project and the identification of a system with which to make the proceedings available online. Although conference proceedings are being published across a number of colleges at Cornell, the project focused on surveying conference proceedings produced by the Life Sciences libraries. To identify a small set of conference proceedings with which to begin the project, a number of Life Sciences librarians and faculty were interviewed. In the end, the proceedings for the Wine Industry Workshop, a wine production conference held annually in Geneva, New York, were chosen for the project.

DSpace

Once the set of proceedings was chosen, a survey of the various digital library platforms available to provide online access to the proceedings was begun. This examination showed that Cornell was experimenting with a wide variety of digital library platforms. These included DLXS, developed by the University of Michigan and used by Cornell to support its Core Historical Literature of Agriculture and Home Economics Archive projects; DPubS, an open source digital library platform developed at Cornell and used to support Project Euclid; FEDORA, developed by the Digital Library Research Group in the Cornell Department of Computer and Information Science in partnership with the University of Virginia; and DSpace, the institutional repository developed by MIT and Hewlett-Packard and put into experimental use at Cornell. (3,4,5,6)

DSpace was chosen for the project for a number of reasons. First, Cornell University Library had recently installed an instance of DSpace and was looking for some projects to experiment with. Second, one of the members of the committee awarding the internal grant that funded this project was an administrator in CUL's IT department and she recommended DSpace. Third, DSpace's hierarchical structure of "communities" and "collections" was viewed as useful for structuring the proceedings. For the project, the proceedings were to be structured in such a way that researchers could search the collection from numerous levels:

- Entire collection of proceedings;
- Conference proceedings produced in one particular college (Agriculture and Life Sciences, for example);
- Proceedings from one particular conference in a single college;
- Single year's set of conference proceedings from a particular conference; or
- Individual proceedings within a single year.

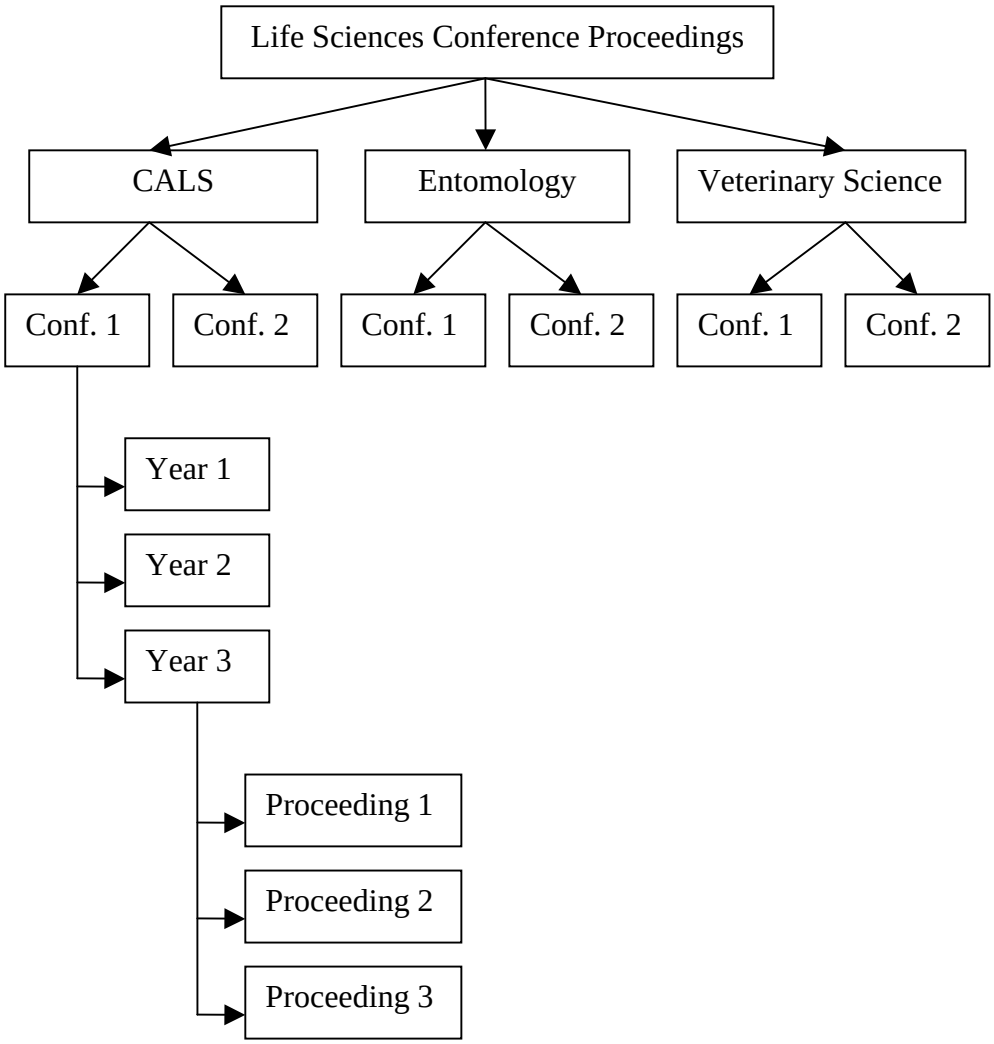

Diagram One: Hierarchical Nature of Conference Proceedings

Fourth, DSpace's metadata creation tool seemed very useful in assisting those adding proceedings to the repository. In the past, metadata librarians at Cornell have had the opportunity to work on digital library projects in which authors of digital information resources also created the metadata associated with them. The librarians have learned that sometimes the authors create less than perfect metadata, which subsequently needs to be cleaned up, and massaged for inclusion in the digital library system. Any tool that enables easier, more efficient creation of metadata is useful.

Gathering and Scanning Proceedings

After a set of conference proceedings and an institutional repository in which to store them were selected, the proceedings themselves were collected. This process involved contacting the administrative staff and the faculty member responsible for the conference; the staff was able to provide the library with both print and digital copies of four years of proceedings. This ad-hoc method of gathering the proceedings worked well for this project, but if the project were to be made a permanent part of the library's programs, a specific workflow would have to be developed for the identification and gathering of proceedings. The program managers would have to work with the library director of public relations and engage in one-to-one marketing to create awareness on campus of the conference proceedings repository. Methods would have to be designed for moving the proceedings from the authors' desktops to those of the librarians administering the project. A system would need to be developed for storing the proceedings in paper or digital format between the time of their receipt and the time they are actually loaded (in the case of digital files) or scanned for

loading (in the case of print copies) into the repository. Scanning proceedings for loading into a repository is not inconsequential; the project managers would have to select a scanning vendor from the vendors available in-house at Cornell and outside the University. Care would have to be taken to ensure that the scanned files received for inclusion in the library were accurate and of good quality. Fortunately, Cornell University Library in general—and Mann Library in particular—has a great deal of experience working with scanning vendors.

Loading Proceedings into DSpace

After the proceedings were collected, they were loaded into DSpace. Through this process, it quickly became evident that DSpace may not have been the best digital library platform/institutional repository to use for the project. First, DSpace's hierarchy proved not to be extensive enough for the purposes of the project. The hierarchy in the initial release of DSpace was limited to just two levels, "communities" and "collections." Although the hierarchy was expanded to three levels with the addition of "sub-communities" in a 2004 release, neither hierarchy was extensive enough for the purposes of this project. At least four levels of hierarchy would be needed to provide useful access to conference proceedings in the life sciences project. While DSpace is based on the Dublin Core metadata scheme, which incorporates hasPart/isPartof refinements to its "relation" element, the use of these refinements to provide access to the various levels of the hierarchy of materials is not the best solution to this problem. DSpace would be more useful if it could be modified to support hierarchies of different sizes depending on the collection in question.

In addition, DSpace's metadata creation form proved to be less useful than was originally thought. It was lengthy and complex; even the metadata librarian working on the project found that using the form to create metadata for the proceedings was tedious. Authors of conference proceedings not skilled or interested in metadata creation would find this form even more of a barrier. It would be more practical to design a metadata creation form that still utilized Dublin Core but was better suited to the needs of local projects. Such a form could be used to create a single metadata file representing metadata for all the objects in a single repository; this file, in turn, could be batch loaded directly into the DSpace repository, bypassing the DSpace metadata creation tool.

On the other hand, it is possible that no metadata creation form could be designed to enable proceedings authors to easily create metadata for their proceedings. A number of current digital library projects that receive content from external authors have needed to assign a project staff member the task of editing the metadata accompanying any content received. This may also need to be done for the proceedings project: a librarian may need to coordinate the creation of metadata for the proceedings upon their receipt. A tool best suited to meet the metadata creation needs of library staff, rather than proceedings authors, may need to be developed. The characteristics and features of a metadata creation tool used by library staff will differ from those of a tool intended for use by proceedings authors external to the library. The process in which metadata is to be created—whether it is to be created by the authors of the proceedings or by librarians after the proceedings are submitted—will need to be determined before any metadata creation tool is developed. The development of a metadata creation tool could mean a refinement to the DSpace tool or the development of an entirely different one.

In this study of DSpace, it was learned that its preservation mechanism (7) — converting the objects within the repository into bitstreams and preserving the bitsreams—has been called into question by a number of preservation professionals. Even though bitstreams may be easier to preserve for long periods of time than are, for example, Microsoft Office documents, the process whereby the objects within the repository are converted to bitstreams and then converted back into objects needs to be carefully evaluated. It is one thing to convert a Microsoft Office document into a bitstream for preservation and storage in DSpace and then convert that bitstream back into a Microsoft Office document. It may be something else entirely to convert a Microsoft Office document into a bitstream for preservation and storage in DSpace and then, in the future, convert that bitstream into a format that does not currently exist.

As consideration has been given to the design of a library service based on the findings of the proceedings project, DSpace's limitations have prompted further exploration of other digital library systems. There are advantages and disadvantages to each of the digital library/ institutional repository options currently available at Cornell. The DSpace designers have gotten the hierarchical design of the system right; they just need to expand it. Another repository that has a built-in hierarchy and is

structured to work with journals is DPubS. Since DPubS is an open source platform, it will improve with time as implementers add new features. In addition, DPubS is already in production mode, so those using the DPubS platform for other projects will be able to rely on the expertise of those currently working with the system. Even though DLXS is also supporting projects that are in production, it does not have a DSpace-like hierarchy built in and seems to work best with monographic materials. FEDORA is an excellent digital library platform, but it only supports a repository back end; the development of a user interface for any FEDORA-based system would involve a major commitment of time and money.

Economic Concerns

In addition to the financial investment required for the development of a FEDORA-based system, similar investments may be required for a number of other pieces of this project. Scanning back runs of print proceedings for inclusion in the repository would not only involve the selection of a vendor but also the design of a business plan to pay for the scanning of the proceedings; scanning costs vary significantly among vendors. In addition, if a library service was developed in response to this project, decisions would have to be made about how these scanning costs would be recovered. The costs could be passed on to the authors of the proceedings or even users of the site. Another cost that would have to be addressed is the cost of proceedings for current conferences. When the repository was demonstrated for an audience at the 2004 Wine Industry Workshop, the question most frequently asked was whether the proceedings should be made available for a fee. Workshop attendees currently pay for the proceedings as part of the conference registration fee. If the proceedings were made freely available via an institutional repository to both workshop attendees and non-attendees alike, the non-attendees would have an unfair advantage, since they would not have already paid for the proceedings through the conference registration fee. On the other hand, one could ask if a charge would still be levied for the proceedings if they were made available online. If that charge is currently being levied only to cover the production and binding costs, then that charge could be eliminated if the proceedings were made available online and not produced in paper form and made available at the conference. The costs of "producing" the proceedings would then fall to the conference attendees, who would need to go to the repository, download the proceedings, and print them out themselves.

Additional Technical Issues

If the system needed to distinguish between those who had attended the conference and had "free" access to the proceedings and those who had not attended and did not have access, different levels of access would need to be built into the system. This might be accomplished with user recognition devices such as user names and passwords, which may not be built into to a particular digital library/institutional library system (it is not built into DSpace). This would be an important design consideration in the development of any system.

A second technical consideration involves the life cycle of conference proceedings. Abstracts for conference papers are sent to the conference conveners before the conference, Microsoft PowerPoint slides are often created in connection with the papers, and the actual papers are submitted to the conference conveners separately. Each conference "proceeding" may consist of these parts; if a conference proceedings repository recorded these separate parts, there would need to be a way to tie them together so a user searching for a proceeding would be able to retrieve all components of the proceeding. Similarly, the repository would have to be structured in such a way that it could store and provide access to various types of content such as Microsoft Word documents, Microsoft PowerPoint slides, and other types of files.

A third technical consideration involves the extent of any collection of conference proceedings. It is one thing to create a repository for a single institution's proceedings, but a much larger effort to create a network of repositories storing proceedings from multiple institutions. This would be a logical extension of any proceedings repository; no single institution sponsors all conference proceedings, so it would be beneficial to network a number of proceedings repositories together to provide shared access to a large number of conference proceedings. Institutions would need to partner with one another to create a shared repository of conference proceedings. In addition, the conference proceedings repositories would need to interface with one another; this would be another important design consideration in the design of a conference proceedings repository. Standards and protocols like the Open Archives Initiative

Protocol for Metadata Harvesting would need to be implemented in each one of the repository systems to make them compatible with one another.

Conclusion

In conclusion, to fully support the three-fold mission of Cornell University, the library needs to provide access to all types of information resources, including grey literature that is not as readily available as other information. One particular type of grey literature that has been identified as difficult to access is conference proceedings. This project explored the use of a particular institutional repository—DSpace—in providing access to these conference proceedings. This project enabled Cornell University librarians to learn about the issues surrounding the use of an institutional repository—or other digital library platform—to provide access to conference proceedings. This project's findings will be used to determine whether or not a service of providing access to conference proceedings via a digital library platform is viable within the Library.

References

1. Dankert, Philip; Gregory Lawrence; Patricia Viele (2003). Grey Literature: A Research Report (2002-2003 CUL Internal Grant). Ithaca, N.Y.: Cornell University Library, p. 1.
2. Weintraub, Irwin (2001). The Role of Grey Literature in the Sciences. http://library.brooklyn.cuny.edu/access/greyliter.htm/
3. DLXS: Digital Library Extension Service. http://www.dlxs.org/
4. DPubS. http://dpubs.org/
5. The Fedora Project: An Open-Source Digital Repository Management System. http://www.fedora.info/
6. DSpace. http://www.dspace.org/
7. Bass, Michael J, et al. (2002). DSpace – A Sustainable Solution forInstitutional Digital Asset Services – Spanning the Information Asset Value Chain: Ingest, Manage, Preserve, Disseminate: Internal Reference Specificiaton: Functionality. Cambridge: Hewlett Packard Company; Massachusetts Institute of Technology, p. 4-5.

https://doi.org/10.26069/greynet-2026-000.613-tgj

Grisemine,
A digital library of grey university literature *

Marie-France Claerebout (France)

Abstract
The progressive change from printed to digital support, induced by the policy of the commercial publishers, is likely to reduce little by little our university library to being a passive relay, and consequently to question our existence in the circuit of the dissemination of scientific information. Our survival depends on our capacity of adaptation. That is why we chose to create without delay our own digital library of grey literature, Grisemine, and to transpose our know-how as information specialists in the necessary tasks, which are indexing, cataloguing and organization of the access to documents. Created at the end of 2001, Grisemine has gradually grown richer by a selection of teaching or research documents: theses and dissertations, courses, communications, scientific reports... in the various disciplines, which our university library covers. The favorable reaction of the contacted authors, as well as the increasing frequentation of our site, (http://bibliotheques.univ-lille1.fr/grisemine), confirm the relevance of our step. It still needs to reach a critical size, a principal guarantee of credibility. We hope to reach that point while developing the co-operation with other actors engaged in the production and/or dissemination of grey university literature.

Introduction

For years, our university library has been used for collecting printed grey literature and communicating it to her readers. Most of those documents (about 60,000) are theses and acquired from national or international exchange among libraries. The others are donated by institutions or by the authors themselves e.g. working papers, conference proceedings, official reports, duplicated lecture notes, etc.
The dissemination of grey literature has become a stake for researchers, in reaction to the price inflation of commercial publications. That is why we can witness an evolution of the scholarly information dissemination, initiated by the authors themselves and helped by the development of the World Wide Web. Before being accepted by well-known journals, articles are quickly proposed to the community on their author's home page, on the institution's web pages, or in some specialized preprint open archives.
Indeed, the increasing possibilities brought about by electronics allow a new way of dissemination for knowledge, the Net. That is the way chosen by commercial publishers; subscribers of most scientific journals can directly read them in their laboratories on their computer screens.

Does this then mean that while knowledge use to be disseminated by means of libraries within the context of printed publications, it will now be disseminated by means of a network in the world of electronic information? Are new technologies going to replace libraries and librarians? We academic libraries, have to concern ourselves with this. Are we becoming less involved in the scholarly publishing process? The response depends on us. Librarians have to demonstrate that taking into account their know-how will help the whole community.

A first step is to continue providing information. Our traditional role still exists, however, documents are both in print and electronic form, and this role is to collect, catalog, preserve, share, and provide easy access to texts. As the physical library is a central part of the university, the digital library ought to be a central part of the "on-line university", the non-for-profit actor offering necessary documentary services as a supplement to educational services.

Besides providing access to more or less expensive e-journals, our website should offer access to grey literature where an important sum of knowledge lies. Our confirmed skills allow us to build by ourselves organized grey literature repositories; I mean true "digital libraries", which we have started to do with Grisemine.

* First published in the GL5 Conference Proceedings, January 2004

GRISEMINE: A FIRST STEP

At the end of 2001, we created Grisemine, a digital library destined to be a mine of grey literature. Grisemine is gradually growing - slowly because of our limited means but safely because of the value of its contents.

Several goals are pursued:
- Giving more visibility to French-written scholarly production
- Bringing our usual readers, scientists and students, wider access to research
- Developing new collections
- Getting new distant readers
- Developing an economical complement to commercial scholarly publishing

Contents on line
In this first step, we only collect French grey literature. A development plan was defined, which aims at a relative weight for each selected discipline and for each selected type of document.

Texts from any country and in any discipline covered by our university are accepted, including the sciences, technology, and the social sciences.
We look for three types of documents produced by universities, research and/or teaching institutes, and public organizations. These are:
- First, research texts which have already been published as theses and dissertations, symposium communications, articles in scholarly journals (journals published by societies, many of which do not yet provide electronic versions of their journals)
- Second, research texts which have not yet been published, such as notes, working papers, and preprints
- Third, teaching manuscripts, such as course outlines, exercises, exams, etc.

Much like a traditional library, Grisemine has to collect, catalog, preserve texts and make each easy to retrieve. We try to answer our readers' current needs, using today's tools. So far we have kept a hybrid approach where most resources are both available in printed form in the physical library and in digital form on the network.

We apply new technologies for making the full text available in electronic format through a search interface on the Internet. A user can:
- Browse the digital library by subject or by discipline
- Search across title, author's name, keywords, table of contents, etc.
- Search inside full text
- Freely print or download texts

So we offer distant printing and navigation inside the text, which are not actually the traditional services of a library.

Scientific and legal aspects
One of the most frequently asked questions is "If some texts are already available on other Websites, why do you not simply provide links to those sites?" This could be one approach, but ours is quite different. As librarians, we feel that our mission is to provide guaranteed, high quality, and secure contents.

Asking for a copy of each text, and building our own library, can seem a heavy task, however, controlling the entire information chain is also a guarantee of durability. We cannot risk loosing precious documents, on account of a broken link.
It is likewise a guarantee of credibility and quality. We can refuse to process a text that would not be considered scientifically sound. For any writer to cooperate with Grisemine, the only requirement is to belong to a known research or teaching institution. So, with just a little more control, we are able to provide rapid dissemination. And, if needed, peer review could easily be envisaged.

Before any copying or dissemination, we ask the copyright holder for rights. It is sad to note, but in the case of previous (even non-commercial) publication, authors all too often loose control of the intellectual property or, at least, are not sure they have kept it.

Technical aspects
We collect French grey literature, in both printed and electronic formats. By now, most of the documents exist in primary digital sources, which helps to provide access in a cheaper way.

For insuring security and fair use of the electronic resources, texts are converted in PDF format, protected against 'copy-paste'. In this way, the distant-printed version matches the appearance of the original one. A system of bookmarks enables the use of the table of contents as a navigation tool inside the text, while the hidden text permits searching inside the full text. With the author's agreement, we are also planning to offer HTML format.

Bibliographic description
Whether the format is printed or electronic, the more documents you get, the more structuring you need. Our search interface on the Internet[1] reflects our tools and methods of bibliographic description.

Each electronic text is linked to metadata tagged in MarcXML. Created by the Library of Congress, the MarcXML format mixes strictness of Marc, the librarian's format, with interoperability offered by XML.

Another guarantee for interoperability is to maintain international standards inside those metadata fields. Each text has a double discipline classification: a local one, useful for searching on the Website, and an international one, the Dewey classification, useful for metadata exchange or pooling.
With the same objectives, which are pertinence of results on the Website and interoperability, we also systematically define normalized keywords alongside terms chosen by the authors. The description of each text also includes the abstract and table of contents. As well as the original quotation, if the text has been loaded from another Website.

GRISEMINE'S FEEDBACK

After two years, results have been quite good. In spite of the small quantity of available texts (about 750 as of November 2003), frequentation of our website is steadily increasing.
Originally, we had to correspond a lot in order to become known. Now, people come more spontaneously to our site; and, these are probably not all French-speaking people. From the beginning of 2003, visitors have come from 74 countries that encourage us to collect and disseminate in other languages, which are too scarce on the Web.

Who are our distant readers and what are they looking for?
As expected, they do not only come from academic institutions. Several firms, very interested in scholarly grey literature, have contacted us asking for texts on specific issues. We have also noticed a great interest in course outlines and teaching materials. While they represent only 11% of the contents, they account for over 30% of Grisemine's retrieval. So, on the last revision of our development plan, we decided to increase the proportion of teaching texts within the digital library.

And what about the authors?
When contacted, authors usually agree to give us their texts - even if they often forget to do so and have to be re-contacted - a few new authors have willingly begun to send their texts. Their trust in the visibility offered by our digital library is a good credibility indicator.

Closer to the authors
A new challenge for librarians will be to take into account the structuring of content. Even if grey literature is heterogeneous in essence, some kinds of documents can be made homogeneous from their creation, e.g. theses, dissertations, and articles. Why should we then ask the author to respect a few rules for structuring documents? Simple, in order to get a better printed version, in much the same way as some publishers and symposium organizers do. The aim is also one of preservation, because

a structured document can be archived in a durable format and allows further modular re-use of the text.

In France, our university is one of the first to engage in the process of structuring theses for electronic diffusion and conservation. Theses will be tagged for XML encoding designed to identify each element of the title page and each part of the thesis. It is with this aim in mind that the library provides authors useful recommendations about the style sheet to use, once they have begun writing.

Another development for university libraries will be the creation of institutional repositories (IR). Indeed, more and more institutions want to keep ownership of information produced by their members and so, they plan to create a digital repository for their entire intellectual output. In this way, libraries can be better service providers.

Together these new aims add to the information provider a more service-oriented role and lead us to a new position in the information cycle of creation, distribution, and use.

In close, collaborating with other actors

Grisemine can be seen as a prototype for other projects of scholarly grey literature dissemination often led by researchers themselves. There still remains the objective to manage vast amounts of electronic information and facilitate retrieval based on cooperation among the different actors. Generally speaking, all not-for-profit actors would probably seek the advantage of cooperation between one another *i.e.* libraries, higher education, researchers, scientific societies, not-for-profit publishers, university presses, etc.

However, first and foremost, document providers should cooperate by exchanging full texts and/or metadata. This would then allow for large digital libraries or large metadata repositories. Grisemine will soon be able to do this by becoming OAI-compliant. This means that it will respect the existing Open Archive Initiative protocol accompanied by still missing normalized semantic rules.

References

[1] http://bibliotheques.univ-lille1.fr/grisemine

https://doi.org/10.26069/greynet-2026-000.613-tgj

Wallops Island Balloon Technology: Can't see the Repository for the Documents *

Andrea Japzon and Nikkia Anderson (United States)

Abstract
Since the Wallop's Balloon Technology documents repository began approximately 9 years ago, the Goddard Library has become increasingly involved in developing digital archiving capabilities. The Library developed the Digital Archiving System (DAS), which is a prototype infrastructure for creating a combined metadata repository that allows metadata for heterogeneous digital objects to be searched with a single search mechanism, and presented in a single results page. With this, the opportunity has been presented to expand the usability of the print repository. The Balloon Technology documents relate only to the specific subject of construction of scientific balloons and at the current time number over 4,300. The documents exist primarily in paper format and are organized according to the accession number. The project is currently at a crossroads where decisions will have to be made regarding the reorganization of the database from many different perspectives. An assessment of the project was conducted to determine future direction. An assessment survey was created using the Knowledge Management Assessment Tool (KMAT) from the American Productivity & Quality Center and from the recommendations that The Scholarly Publishing & Academic Resources Coalition (SPARC) put forth in "The Case for Institutional Repositories: A SPARC Position Paper. Survey participants agreed to move forward with project by scanning the documents, mapping existing database records to the current metadata elements, seeking copyright permissions, and forming a joint committee of balloon scientists and librarians. All have agreed on the importance of digitizing this collection to the balloon science community. Further, it was agreed that once complete, the addition of the balloon documents to the DAS (an institutional repository) could serve as a model for other NASA and/or government projects trying to organize, collect and preserve specialized knowledge that manifests largely in grey literature.

History of the Library and the Database

The Balloon Technology Library and Database was created to provide a single repository for scientific ballooning literature and data. Grey and commercially published literatures comprise the collection of technical reports, working papers, proceedings, and journal articles. The primary impetus for creating the library and database was the fact that much of the literature relating to balloon technology is grey literature and therefore not easily found, let alone found in one place. The Balloon Technology documents, currently numbering over 4,300, relate specifically to the construction of scientific balloons. The subject matter covered includes balloon theory, balloon physics, design, operations, performance, facilities, testing, materials, fabrication, quality control, failure analysis and history [1].

The Balloon Technology collection project started in November of 1993 continues to exist today with the sustained involvement of Wallops Flight Facility (WFF) balloon scientists and Goddard Space Flight Center (GSFC) librarians. WFF is a part of GSFC but located three hours south of Greenbelt, Maryland in Wallops Island, Virginia. The WFF mission includes the performance of research in the areas of sub-orbital and small orbital payloads. Balloon science experiments are a significant part of this mission. From the beginning of the project, the intention of the Balloon Program Office (BPO) at WFF was to start a library that would become the World's Archival Center for Scientific Balloon Technology [2]. To that end, project participants set out to and accomplished the major goals of collecting to the greatest number possible of existing balloon technology documents and creating a searchable database. With these accomplishments in hand, now is the time to increase the availability of this work to balloon science researchers.

A great debt is owed to Jim Winker, balloon scientist and de facto librarian, for performing the yeoman's duty in reaching the goals set out by the BPO. In the field of librarianship, perhaps the greatest value placed on information is its authority. In the

* First published in the GL6 Conference Proceedings, January 2005

creation of this collection, Jim Winker's life experiences provide that authority. He possesses extensive historical and technical knowledge of scientific ballooning. His knowledge has developed through "a long continuous interest and participation in the field of scientific ballooning and interaction with the many present and past commercial and government organizations" [3]. The knowledge and experience of individuals within a specialized field is what sustains the use and value of grey literature within the field. Jim Winker's large-scale interaction with balloon literature clearly exemplifies this.

Mr. Winker visited over 60 sites in his effort to find and select documents for the library [4]. He started locally with his own collection, the Raven Industry library, private collections in Sioux Falls, South Dakota and the holdings of Wallops. His search expanded to include the information repositories of the National Technical Information Services (NTIS), Defense Documentation Center (DDC), Association of Balloon and Airship Constructors (ABAC), and American Institute of Aeronautics and Astronautics (AIAA). He explored organizations that use balloons currently or have used them in the past. These organizations included research centers, manufactures, scientists, research sponsors, and operations centers. In his effort to be comprehensive, his search included an "everything else" category as well. He evaluated the resources of libraries, museums, individuals, published works, and international resources [1].

With the care of an archivist, Mr. Winker researched preservation quality paper, durable binding, and copyright implications [5]. While he realized the selection and preservation of the documents was of fundamental importance, he also realized that "the database may well be the most important part of the project. Without it, the library would have a far more limited benefit" [1]. The database would promote the existence of the collection and ultimately provide access to the documents. Working with Janet Ormes, GSFC Library Head, Mr. Winker, selected software to support the database and created guidelines for cataloging the documents. The COSATI Standard for Descriptive Cataloging of Government Scientific and Technical Reports was relied upon for the creation of the rules of entry [6]. For each document, an entry was created for the database using a variety of descriptors intended to aid in the process of information discovery.

The askSam database was chosen in 1994 and began production with the database successfully organized and defined [7]. At this same time, database search methods were explored and decided upon. Progressive for the year, both controlled vocabulary and natural language searching were made possible [8]. In the ten years that the documents were collected and that the database grew in significance for the ballooning science community, technology was developing that would further enhance the work of Mr. Winker.

Current Status of the Documents

Since the Wallops' Balloon Technology Library began, the Goddard Library has become increasingly involved in developing digital archiving capabilities. The Library developed the Digital Archiving System (DAS), a prototypical infrastructure for creating a combined metadata repository allowing for heterogeneous digital objects to be searched with a single search mechanism and presented in a single results page. The DAS is an institutional repository of scientific and technical information including selected Goddard web sites, videos, images and documents. With this, the opportunity has been presented to expand the usability and accessibility of the balloon technology library.

The documents exist primarily in paper format and are organized according to the accession number. The documents are housed as a separate collection at the Wallops Island Technical Library. The database is accessible to all from the library's website. NASA researchers can gain access to the documents by either using the library at Wallops or having the documents scanned and emailed to them. Researchers outside of NASA can make requests from Wallops with some restrictions placed on their access.

Since January of 2004, some balloon technology documents have been scanned as a pilot project into a MySQL database as PDF files. The askSam records as contained in a WAIS database link to the PDF files; however, the DAS, which the MySQL database supports, does not relate to the askSam database. The askSam software is no longer being developed. It has limited capabilities for interacting with the World Wide Web and the searching environment. Currently we have two separate search interfaces, one is a meta-search in WAIS and the other Autonomy, which indexes the documents and has full text search capabilities [9,10]. WAIS and Autonomy can't be configured to work together to provide meta- and full text searching. The DAS can do both by using

MySQL and Lucene. Not to mention, the DAS allows you to search, archive, and preserve documents for future use. By incorporating the balloon technology documents into the DAS, the balloon science community will have increased visibility and access through meta-, full-text searching and retrieval options to include the full-text.

The project is currently at a crossroads where decisions will have to be made regarding the incorporation of the database into the DAS and the accessibility of the documents on many different levels and perspectives. The authors of this paper turned to the community involved in the creation, the delivery and the use of the library and database to determine the future direction of the project.

Knowledge Management Assessment

An assessment survey based on the Knowledge Management Assessment Tool (KMAT) from the American Productivity & Quality Center was administered to nine individuals, representing different perspectives on the project [11]. Further, these same individuals assessed the project based on the recommendations that The Scholarly Publishing & Academic Resources Coalition (SPARC) put forth in "The Case for Institutional Repositories: A SPARC Position Paper [12]. **See Appendix A** for the complete survey assessment. A meeting was held on August 12, 2004 with survey participants and Balloon Program Office (BPO) representatives in attendance. Issues brought forth by the assessment created the agenda for the meeting.

Given that the work of the Balloon Technology collection is a small slice of a greater organizational structure, survey participants were ask to be cognizant of this while responding to the KMAT as knowledge management is the function of an organization. The authors felt that the current status of the project presented an instance of the process of knowledge management in play. Mr. Winker is retiring for the second time in January 2005 and will no longer be working on the project. As stated earlier, he has performed the vast majority of document collection and cataloging. The authors wanted to examine how well the knowledge that Mr. Winker possesses is being transferred to the individuals remaining on the project.

The greatest benefit of the knowledge management assessment indicated an important difference in perspective among survey participants. Two perspectives emerged: content and infrastructure. Individuals with a content perspective included contributors and users of the information contained within the collection. Individuals with an infrastructure perspective included systems, web, library, and archival staff that primarily work to make the collection accessible to users of the collection.

At the assessment meeting, both perspectives were defined and the importance of each established. Certain points of confusion regarding current communication between all project participants were made clear. Traditionally, the individual who creates information and/or collects the information is not the same individual that catalogs, stores, and distributes the information. For ten years, Mr. Winker was performing most tasks related to the collecting, cataloging, and searching of the documents, not to mention creating literature as well. Enter the possibility of the DAS and open access archives and the flow of knowledge from old processes to new processes does not transfer without some adjustment. The second part of the assessment devised from the SPARC paper complimented the first as it helped the discussion participants describe the processes under consideration for change.

Can't see the Repository for the Documents

For the purpose of this paper, the definition of an institutional repository as put forth by the Scholarly Publishing & Academic Resources Coalition (SPARC) will be used. Institutional repositories are digital collections that capture and preserve the intellectual output of a particular community. Repositories are intended to expand access and retain control over the scholarship produced by that community. Further, the repository has the potential to contribute to the community or the institution it is a part of by providing tangible indicators of the community's quality through the demonstration of the scientific value of its research activities [12].

Being a part of the DAS can in and of itself increase the visibility, status and public value of the balloon science community. Money, time, and technology aside, the documents themselves create the biggest obstacle to creating an institutional repository at Goddard as put forth by SPARC. The SPARC paper further defines an institutional repository as "open and interoperable." To promote interoperability and open access, institutional repositories provide access to the broader research community through no or low barrier access. Either by providing a search mechanism

with indexing or by maintaining and exposing metadata to be harvested by other institutions, interoperability is gained among institutions [12].

While the BPO is interested in sharing the database of records in the manner described above, the office is not interested in sharing the documents in a global capacity. The proprietary nature of some of the documents and the competitive environment in which they were created prohibit the collection of documents as a whole from being accessed outside NASA IP ranges. Small Business Innovation Research (SBIR) documents are an example as they are restricted for five years after their release. The BPO representatives want to keep the collection primarily for balloon scientist with a relationship to NASA research, hence limiting access to greater research community. It is extremely likely that other Goddard projects will have similar concerns or restrictions regarding their documents. Fortunately, the technology exists to allow different levels of access to different types of information allowing Goddard to interoperate fully in at the metadata level.

If the database records are to be open and interoperable, the existing data fields from askSam will have to be mapped to the metadata elements of the Goddard Core. The Goddard Core is a metadata element set and is employed in the DAS single search mechanism. Metadata is information that describes a digital object; like a library catalog card for a digital object. The Goddard Core Metadata set contains 24 elements that describe project-related objects of interest to GSFC. It extends ISO 15836 (Dublin Core) Standard for Descriptive Metadata for Electronic Resources while using Open Archives Initiative standard protocols that can interoperate with other systems.

When Mr. Winker was devising his catalog entries, he was working in a contained environment. There was no need to consider how the descriptive fields he employed related to a larger information structure. The technology did not exist for interoperability to be an issue. The Goddard Core establishes guidelines for organizing documents and other objects in a way that is both meaningful and accessible across GSFC, within NASA and in the open access environment.

The mapping will involve more than a programming solution, as the fields used in askSam do not match directly or in some cases not at all to the Goddard Core elements. Without ISO standards in place, definitions of descriptors were not nearly as controlled. The Goddard Core allows for project profiles so if need be the 24 elements can be expanded to accommodate the particularities of a document or object collection. The mapping of askSam fields to Goddard Core is a key area for the collaboration from both content and infrastructure perspectives.

Copyright presents another problem in bringing an already existing print library and database into the realm of the online repository. To date, there are 4,330 records in the database of which 847 are indicated as copyrighted. That is roughly 20%. A copyright campaign will need to be launched on behalf of the repository under the auspice of the BPO. This will take a significant amount of time and agreement is not guaranteed. It was agreed that copyrighted documents will begin to populate the DAS and as permissions are gained documents will be added with the understanding that all documents are not likely to be included in the online repository.

As mentioned previously, Mr. Winker is retiring. His knowledge is the current certification and collection process. While encouraged by the ease of collecting new submissions to the collection through online technologies, the users of the collection and the BPO representatives were concerned about the quality that self-selection produces as well as the possibility for submitting items out of scope. Within the balloon science community, there is considerable interest in expanding the collection to include literature related to the experiments conducted using the balloons. Further, online submissions would not replace the need to produce an archival paper copy for the balloon technology library. The meeting participants decided that an official joint committee of Goddard librarians and BPO staff should be sanctioned to establish a certification process and to address any further decisions that need be made regarding the repository. Additionally, the joint committee will facilitate and encourage communication between infrastructure and content perspectives.

Conclusion

As the survey participants sat in the room where the Balloon Technology library is housed, a collective realization fell over the group at the conclusion of our assessment discussion. The ways in which people seek and interact with information are dynamic and online. The print collection resting on the shelves seemed anachronistic given our discussion. The BPO and the Goddard Library agreed to move forward with scanning the documents, mapping the records, seeking copyright permissions, and forming a joint committee. All have agreed on the importance of digitizing this collection to the

balloon science community. Further, it was agreed that once complete, the addition of the balloon documents to the DAS institutional repository could serve as a model for other NASA and/or government projects trying to organize, collect and preserve specialized knowledge that manifests largely in grey literature.

Currently, funds are being pursued to make the transition from a library and a database to an institutional repository a reality at Goddard. Given the nature of the types of information produced at Goddard, any institutional repository created at Goddard or a like environment will not be able to participate fully in open archive initiatives. In our case, the DAS works as an institutional repository for Goddard and to a lesser degree beyond the walls of NASA IP addresses: metadata is acceptable for web harvesting but most documents would not be freely available.

The DAS is still in beta testing and contains over 90,700 web pages, 900 images, and 400 videos. The Balloon Technology document collection will be the first document collection to be included in the DAS. The choice of this collection has proven fortunate in that it exposed many of the issues we would likely encounter when considering the addition of other collections.

References

1. Winker, J. A. (1994). A Balloon Technology Library Database. Paper from the 32nd Aerospace Sciences Meeting & Exhibit. Reno, NV, January 10-13, 1994.
2. Winker, J. A. (1993). Balloon Technology Library Project Monthly Progress Report No. 1. December 3, 1993.
3. Winker, J. A. (1993). Balloon Technology Library: Discussion topics for initial project meeting of 23 November 1993. November 18, 1993.
4. Winker, J. A. (1998). Balloon Technology Library Project Monthly Progress Report No. 45. January 2, 1998.
5. Winker, J. A. (1994). Balloon Technology Library Project Monthly Progress Report No. 3. February 1, 1994.
6. Suggested Guidelines. (1994). Condensed from Guidelines for Descriptive Cataloging of Reports: a Revision of COSATI Standard for Descriptive Cataloging of Government Scientific and Technical Reports.
7. Winker, J. A. (1994). Balloon Technology Library Project Monthly Progress Report No. 5. April 2, 1994.
8. Winker, J. A. (1994). Balloon Technology Library Project Monthly Progress Report No. 6. May 4, 1994.
9. Balloon Technology Database. (2004). <http://library.gsfc.nasa.gov/Databases/Balloon/balloon1.html> (October 26, 2004).
10. Search Goddard Library. (2004) <http://library.gsfc.nasa.gov/SiteSearch/premium/premiumquery.html> (October 26, 2004).
11. American Productivity & Quality Center. (2001) The Knowledge Management Assessment Tool (KMAT). <http://www.kwork.org/White_Papers/KMAT_BOK_DOC.pdf> (October 26, 2004).
12. Crow, Raym. (2002). The Case for Institutional Repositories: A SPARC Position Paper. <http://www.arl.org/sparc/IR/ir.html> (October 26, 2004).

Appendix A

Part I

The KMAT was used as presented at http://www.kwork.org/White_Papers/KMAT_BOK_DOC.pdf with the following questions removed: P3, L2, M1, and M3. Reason being the project is produced by a government organization and therefore will not be sold or marketed for profit.

Part II

1. What purpose does an Institutional Repository serve?
2. Policy consideration:
   a. What is the current copyright policy? Does it need to be improved?
   b. How are restricted documents handled? Could this be done differently?
   c. With regards to accessibility, do we need a policy that differentiates between internal and external customers?
   d. What is the current certification process, the process that assures the quality of the documents added to the repository? How will this change once Mr. Winker leaves the project?
   e. When you consider the future of the repository, will we need a formalized accession policy?
3. Does the repository embody the institutional quality of
   a. The Balloon Technology Program YES □ NO □
   b. NASA YES □ NO □

c. If No to either "a" or "b", please explain what needs to change for this to be so?

4. Does the repository have formal or official recognition in the Balloon Technology Community? Please explain why or why not.
5. On a scale of 1 to 5, please indicate how important you think it is that the Balloon Document Repository reflects the following. 1 being not at all important and 5 being very important.
   a. Institutionally Offered _____
   b. Scholarly _____
   c. Cumulative and Perpetual _____
   d. Open & Interoperable _____
6. Please indicate by checking which of the following you think should be reflected in the metadata? Jim Winker's Metadata Elements have been mapped to the corresponding Element in the Goddard Core.

Jim Winker's Metadata Elements Used	Goddard Core Metadata Elements Used	Please Check X
Access Number	Identifier.Persistent (Auto)	
Title	Title	
Author	Creator.Employee	
Responsible Organization	Creator.Organization	
Funding Organization	Contributor.Organization	
	Creator.RecordCreator (Default)	
Date	Date	
Notes		
	Date.Current	
	Date.Created (Auto)	
	Date.RecordCreated (Auto)	
Report Number	ID	
Contract Number	Contributor.Contract	
Description		
Subject Terms	Subject.Uncontrolled	
Content	Description	
Available From		
	Contributor.Code	
	Type (Default)	
	Content.Type	
	Format (Auto)	
	Subject.Discipline	
	Identifier.URL (Auto)	
	Language (Default)	
	Rights (Default)	
	Audience (Default)	

7. Please describe any other elements that you think should be added?
8. Please feel free to make any additional comments regarding the Balloon Technology Documents?

https://doi.org/10.26069/greynet-2026-000.613-tgj

Sharing Grey literature by using OA-x[1*]

Elly Dijk (Netherlands)

Abstract

As part of the Dutch DARE (Digital Academic REpositories) programme, NIWI-KNAW is participating in various projects to enlarge open access to Dutch scientific output including Grey Literature. The Open Source web technology that will be used for harvesting is based on i-Tor, Tools and technology for Open Repositories, developed by NIWI (the Netherlands Institute for Scientific Information Services).

This article deals with the initiative that NIWI-KNAW has taken for developing the OA-x protocol, a modular extension of the OAI protocol, OAI-PMH (Protocol for Metadata Harvesting). We shall discuss the advantages of OA-x and for what projects it will be used. In the protocols of the Open Archives Initiative (OAI) currently in use, information is shared by providing metadata of digital files (data providing) that can be read in by someone else (data harvesting). A URL is used to refer to an object in an external site (often a repository). In certain cases, one needs to go further than just sharing metadata. Certainly in the case of so-called collaboratories, it should be possible to transfer the objects themselves from an external website (or repository) to one's own site. And conversely, it should be possible to upload objects to an external site. Even if only the browsing of objects is required, it is necessary to get to the original document in order to be able to index it.

The OA-x project has been set up to enable researchers and administrators of (digital) archives to be able to unlock, edit, supplement, combine and archive metadata and data (objects) in digital repositories. A protocol for harvesting and uploading objects has been developed in this project. There are also several implementations available: OA-x within a CMS, OA-x as extendable OAI data and server provider, and OA-x as repository filler. We have opted for similar names of verbs as are used in OAI-PMH.

The advantages for authors and administrators of (digital) archives are great. It is possible to place articles or other publications on one's own website as full text and easily export them to a repository such as an institute repository. It is also possible to use OA-x to upload publications to electronic journals (e.g. Studies in Mycology) or to a central address where a grey publication will be produced.

With the aid of the i-Tor technology, it was already possible to index PDFs on one's own website as full text and make them searchable via Google. Thanks to OA-x, it is now possible to index PDFs (or other text files) on external sites as full text too.

In collaboratories, it is not uncommon to use collections of images that are split over various sites. With an OA-x implementation, it is possible to make a collection of thumbnails of the images in these distributed collections in one place (as if the images were collected on one site only). An example of a collaboratory is E-laborate, a virtual joint venture in the alpha and gamma disciplines. OA-x is used by E-laborate to upload datasets to subject-based repositories.

OA-x can also be used to make an (national) electronic depot. Objects can be sent to such a depot or archive from institutional repositories. OA-x makes it possible to not only send the object but also multiple datasets. It is even more important that information about the technical data (e.g. in what version of PDF the object was created) can be sent along with it. These data are of essential importance to a depot because they can be used to see whether an object (and the format of the object) has remained unchanged.

Introduction

As part of the Dutch DARE (Digital Academic Repositories)[2] programme, the Netherlands Institute for Scientific Information Services (NIWI)[3] – an institute of the Royal Netherlands Academy for Arts and Sciences (KNAW)[4] – is participating in various projects to enhance open access to Dutch scientific output, including grey literature. DARE is a joint initiative by the Dutch universities to make their research data accessible in digital form. As well as KNAW-NIWI, its other participants include the Netherlands Organisation for Scientific Research (NWO) and, from the conservation

* First published in the GL6 Conference Proceedings, January 2005

perspective, the National Library (KB). The programme began in 2003 and will last until 2006. DARE is being co-ordinated by SURF, an ICT partnership set up by the Dutch universities.

The original aim of DARE in 2003 was to establish institutional repositories of academic output at all Dutch universities and KNAW and NWO institutes. These would contain working papers and pre-prints, dissertations and theses, research reports, datasets, contributions to congresses and multimedia presentations. These repositories form a distributed network: they can be searched either collectively or individually.

The web technology developed by NIWI for use as a content management system and service provider for DAREnet[5], the DARE website, is called i-Tor[6] (Tools and Technology for Open Repositories). i-Tor has also contributed towards a number of DARE projects at individual universities, as well as to developing a service provider to harvest digital material from the universities and a data provider for the KNAW repositories and others. These results were presented at a meeting in January 2004, after all the universities and a number of KNAW and NWO institutes have set up at least one repository which could searched through DAREnet. These repositories contain a huge range of different material: publications, including grey literature; films; audio fragments; and so on. In total, the repositories now contain more than 20,000 digital files.

The aim of the DARE community in 2004 is to add to these repositories, primarily by increasing scholars' involvement in the programme. Also in 2004, NIWI began the OA-x project. Financed by DARE, this has been established to enable researchers and the managers of digital and other archives to retrieve, edit, add, combine and archive both metadata and objects from digital repositories. As part of the project, a new protocol for the harvesting and uploading of objects has been developed. OA-x, which is also available as a plug-in for i-Tor, is open source technology and can be regarded as a modular extension of the Open Archives Initiative Protocol for Metadata Harvesting (OAI/PMH).[7]

This paper looks first at the i-Tor open source web technology, which formed the basis for the development of OA-x. The four areas covered by i-Tor – content management, archives, collaboratories (digital workplaces) and e-publishing – are all addressed. We then provide details of the OA-x project, describing its background, the OA-x protocol and possible applications. The paper then addresses DIDL and METS, before finally mentioning those projects currently using OA-x.

i-Tor: a new open source web technology[8]

The i-Tor web technology is being developed by NIWI, in collaboration with various universities and KNAW institutes in the Netherlands as well similar bodies in Germany and Belgium. i-Tor can be regarded as a toolbox for use when creating a website, collaboratory, information portal, repository, database retrieval system, and so on.

Content management system (CMS)

i-Tor was developed originally as a content management system (CMS) to upgrade the NIWI website. The old site consisted of static HTML pages maintained by webmasters. Thanks to i-Tor, it is possible for non-specialist staff at the institute and elsewhere to keep their own information up to date. Full-text searches are possible in all data, be it web texts, PDF documents or database records. And the content of databases is accessible by search engines like Google. i-Tor is used in the construction of internet, intranet and extranet websites. Access – reading and writing rights – is governed by permissions. In terms of such aspects as layout, i-Tor can be adapted to suit the user organisation. And it is flexible – for example, there are no coercive limitations in workflow.

This CMS forms the basis for further development of the other areas covered by i-Tor. In partnership with various NIWI departments, the development team is enhancing and simplifying the CMS to ease database retrieval and improve the search functionality. These updates will be completed within a few months. As well as NIWI, about 30 other institutes and projects – both national and international – have now (October 2004) constructed their own websites with the aid of i-Tor technology.[9]

Archives
NIWI is also developing i-Tor in other areas. Its Department of History, for example, is a participant in the European Visual Archives (EVA) project.[10] This is working on the searchability of archives containing digital images, which generated a request to develop i-Tor into an OAI service provider for retrieving available metadata and an OAI data provider, making data available in open archive form. The OA-x project described below also falls into this i-Tor category. As well as the EVA project, the DAREnet mentioned earlier uses the OAI service provider and data provider. And NIWI is also working with the German Centre for Polar and Marine Research[11] at the Alfred Wegener Institute to create a kind of large-scale service provider.

Collaboratory
The word "collaboratory" is a combination of "collaboration" and "laboratory". i-Tor allows the creation of a virtual research space for academics working at different physical locations. That space is the collaboratory. It can hold large amounts of information – texts, audio, raw data, databases, video and so on – in various formats (including PDF and XML) for sharing in a straightforward and consistent way.

Functions like authorisations for reading and writing rights, weblogging, RSS feeds, and discussion lists and version management are essential to a collaboratory. All are available in i-Tor, or are being developed for it.

One example of such a collaboratory is E-laborate[12], a partnership in the humanities and social sciences. The aims of this project centre on opening up opportunities to share and collaborate on textual material and datasets. The latter is the subject of a subproject, X-Past. As a test case for the Textual Material subproject, the E-laborate Steering Group has chosen the Dutch historical and cultural journal *Vaderlandsche Letteroefeningen*. This is being used to develop electronic tools for co-operation related to texts and textual material. Those tools will also be applicable to other textual material.

E-publishing
Other i-Tor functionalities are being developed in the area of e-publishing. At the request of another KNAW institute, the Fungal Biodiversity Centre (CBS), NIWI has created a publication tool for the production of a digital journal – in this case, *Studies in Mycology*.[13]

As part of the DARE programme, a "similarities" function has been developed for three Dutch universities. This is actually a plagiarism scanner, which can search students' dissertations for stolen work.[14] Another function, still under development, will enable peer review.

The OA-x project

Background
The OA-x project was established to enable researchers and the managers of digital and other archives – including those containing grey literature – to retrieve, edit, add, combine and archive both metadata and data (objects) from digital repositories.

Under the standard Open Archives Initiative protocols currently in use (OAI-PMH), information is shared by supplying metadata from digital files (data providing) in a form, which can be read by others (data harvesting). A reference may then be provided, through a URL, to an external site – often a repository.

In certain cases, however, users wish to go further that simply sharing metadata. Certainly within the context of a collaboratory, it is preferable that the objects themselves be transferable from an external website or repository to the user's own site. And, conversely, users want to be able to upload objects to an external site.

NIWI-KNAW has conducted the OA-x project with the help of a subsidy from DARE. The project was designed to produce a protocol for the harvesting and uploading of objects. At the same time it also addressed the digital durability of the objects stored. For objects to be retained and accessible in the long term, it is not enough to be able to harvest the metadata alone. The objects themselves also have to be harvestable. Yet the current OAI-PMH is inadequate for that. OA-x, on the other hand, makes it

possible to transfer metadata together with the associated objects. And it allows the user to check that the metadata or the object or the combination of the two has been received correctly.

The protocol

In developing OA-x, it was decided that it should be easy to implement for existing OAI-compliant repositories and that everything should be done in as transparent a way as possible, building upon what has already been achieved with OAI-PMH and i-Tor. So OA-x has been designed as a modular extension of OAI. Developed as a plug-in for i-Tor, amongst other forms, OA-x, is open source technology. Some of its features have been implemented within i-Tor, thus creating a proof of concept with which the specific properties of OA-x can be demonstrated.

OA-x has been developed as an extension of the OAI/PMH protocol.[15] It has also been decided to give its commands names similar to those in OAI/PMH.

The following are the most important commands.

- *GetObject(s). Used to harvest objects. The plural GetObjects command has been created for bulk processing.*
- *PutObject(s). Used to upload objects to an external repository. The plural PutObjects command has been created for bulk processing.*
- CheckSum. A unique number, generated using a special algorithm, which can be assigned to an object.

The CheckSum command works as follows.

- Data provider calculates checksum and data.
- Service provider harvests data.
- Service provider requests the relevant checksum from the data provider.
- Service provider also calculates the checksum of the harvested data.
- If the checksums match, everything is in order. In not, there is something wrong with the harvested data.

NB. OAI itself has no checksum mechanism.

The similarities and differences between OAI-PMH and OA-x are shown in the table below.

OAI-PMH	OA-x
Identify	Identify
ListRecords	ListRecords
ListSets	ListSets
ListIdentifiers	ListIdentifiers
ListMetadataFormats	ListMetadataFormats
GetRecord	GetRecord
	GetObject
	GetObjects
	PutObject
	PutObjects

Logically, the OA-x object closely resembles an OAI record. But they differ in the following respects:

- The header with identifier and datastamp.
- The metadata (DC or any other pattern).
- The body (a PDF, image or any other form of binary).

Clearly, this latter point is where OA-x represents an enhancement. An example of an OA-x object is given in *Appendix 1*.

OA-x applications

OA-x considerably broadens the possibilities available to users, thus significantly boosting the potential for virtual collaboration. The benefits for authors and the managers of digital and other archives are enormous.

Some important applications of OA-x are listed below:

- **Use of the GetObject command**

- Within i-Tor, it was already possible for users to provide the PDFs on their own site with full-text indexing to make them searchable by Google. Thanks to OA-x, the same can now be done for PDFs or other text files on external sites.
- Collaboratories often make use of collections of images, which are spread across several sites. An OA-x implementation makes it possible to produce a collection of thumbnails from these distributed collections for placement on one site. This makes it appear that all the images are in one place.
- OA-x can be used to upload datasets to subject-based repositories.
- OA-x can also be used to integrate full-text objects.
- Analytical tools – for grammatical or textual analysis, for example – can be used even when documents are distributed across several sites.
- At large academic institutes, news bulletins are often generated in different places. The OA-x protocol can be used to bring together these bulletins, or parts of them, at a central point.
- Establishing relationships between objects – for example, by using tools such as the "similarities" function mentioned above – is easier with OA-x than in previous protocols.

- **Use of the PutObject command**

- It is easy to upload articles and other academic information sources to the repository of another institution.
- An export from an existing application or database to an external repository can be generated. At present, this is only possible between i-Tor sites.
- Users can place full-text articles or other publications on their own website and at the same time easily create an export to a repository – for example, that of their institute.
- OA-x can also be used to upload publications to electronic journals or to a central address where a grey publication is produced.

- **Use of CheckSum**

- A checksum is a unique number, generated using a special algorithm, which can be assigned to an object. To check that an archive holds the original object – which is important when working in a collaboratory environment, for example – such a number is vital. Even the tiniest adjustment to the data will change the checksum irrevocably and thus alert the researcher.
- Naturally, such a checksum can be extremely useful in such areas as version management and the verification of electronic publications.

OA-x and the electronic depot: a potential special application

OA-x could contribute to the creation of a national electronic depot or archive, using the PutObject command to submit objects drawn from institutional repositories. Not only can the object be sent, but also multiple metadata sets. Even more importantly, technical information – such as which version of PDF the object was generated in – is sent as well. These details can be vitally important to the depot, since they are used to check whether the object or its format has been changed. This is done using CheckSum. Transferring the object to a different environment, such as a new version of PDF, will automatically generate a different CheckSum. If an institution wants to retrieve objects – its own or other people's – from the electronic depot, it can use the GetObject command.

DIDL, METS and OA-x[16]

There is a long-standing wish to be able to able to retrieve actual objects, not just their metadata. And two standards have already been developed to do this: DIDL and METS.

MPEG 21 DIDL (Digital Item Declaration Language)[17] is a metadata format for any digital object and contains all the elements needed to harvest various objects automatically. But it does require agreements to be reached about a special metadata format for the dissemination of a resource. One solution to this limitation would be to make that dissemination part of the protocol. The OAI Working Group could rule on that. In consultation with service and data providers, it would then decide what such

an extension should look like. But it is going to be some time before a consensus is reached on this issue.

Thus far it has proven practically and, in particular, organisationally difficult to make repositories MPEG 21 DIDL-compliant. For it to be usable in developing an electronic depot, the protocol would at the very least have to be enhanced to include the PutRecord command.

Another possibility is to use METS (Metadata Encoding and Transmission Standard)[18]. This is a metadata model in which it is possible to include comprehensive information about a particular resource. In terms of function, METS is comparable with the Dublin Core: it is intended as a standard for data exchange. But METS does not state which protocol is used to share that data. Any protocol can be used: OAI, SOAP and so on. METS' users include the US Library of Congress and its scope in terms of resources is reasonably broad, encompassing text, audio, video, photography and so on, but not datasets.

Since you can use it as XML, METS combines excellently with OAI. It is relatively easy to send an METS XML record using OAI. However, METS is no OA-x. It is not itself responsible for the sharing of resources, any more than DIDL is. Both METS and DIDL are metadata standards in which you can *either* provide links to the resources *or* incorporate the resources themselves in their entirety. The mechanism, which actually ensures that the resource reaches the user, is not specific in either standard. The two possible scenarios can be summed up as followed.

1. The METS/DIDL description contains a link to the resource, and this must be followed in order to download the actual resource.
2. The METS/DIDL description contains the actual resource, which first has to be "cut" from that description and saved separately before it can used.

In this respect, OA-x goes further. Whereas METS and DIDL contain either the resource or a link to it, so that you always receive the resource through OAI if you opt for them, OA-x actually draws a distinction between harvesting the metadata and harvesting the resource. To do this it has added separate commands (GetObject) to the OAI protocol, which METS and DIDL have not.

Current use of OA-x

Within the DARE community, the development of OA-x has prompted a debate as to whether it is actually going to be used in DARE. Contact has been established with the OAI community and the various options – OA-x, DIDL and METS – are being considered.

As mentioned earlier, OA-x has already found a number of applications. The E-laborate collaboratory, a partnership in the humanities and social sciences, is using it to update datasets into subject-based repositories as part of the X-Past subproject. The Bibliography of Dutch Language and Literature (BTNL) is also using OA-x, to integrate full-text objects. A third application is the upload of publications to electronic journals, as is happening at *Studies in Mycology*.

OA-x has been designed as a modular extension of OAI. Some of its features have been implemented within i-Tor (Tools and Technology for Open Repositories), thus creating a proof of concept with which the specific properties of OA-x can be demonstrated.

Full details of OA-x can be found on the i-Tor website, http://www.i-tor.org/oa_x, together with comprehensive descriptions of DIDL and METS.

Appendix 1: Example of an OA-x object (The PDF object itself has been shortened in order to avoid a lengthy example)

```
<?xml version="1.0" encoding="UTF-8" ?>
<OA-x xmlns="http://oax1.cq2.org/OAX/0.1/"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://oax1.cq2.org/OAX/0.1/ http://oax1.cq2.org/OAX/0.1/OA-
x.xsd">
<responseDate>2004-02-06T15:47:02Z</responseDate>
<request verb="GetObject"
identifier="bob">http://oax1.cq2.org/nl/story4test</request>
- <GetObject>
  - <object>
    - <header>
        <identifier>bob</identifier>
        <datestamp>2004-02-06T15:47:02Z</datestamp>
      </header>
    - <metadata>
      - <oai_dc:dc
              xmlns:dc="http://purl.org/dc/elements/1.1/"
              xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/"
              xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
              xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/oai_dc/
http://www.openarchives.org/OAI/2.0/oai_dc.xsd">
              <dc:creator>thijs</dc:creator>
              <dc:type>pdf</dc:type>
              <dc:title>bob</dc:title>
              <dc:date>2004-02-06T15:31:50Z</dc:date>
              <dc:contributor>thijs</dc:contributor>
              <dc:format>application/pdf</dc:format>
        </oai_dc:dc>
      </metadata>
    - <data>
      - <itor:itor
              xmlns:itor="http://oax1.cq2.org/"
              xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
              xsi:schemaLocation="http://oax1.cq2.org/OAX/0.1/itordata.xsd">
              <itor:originalfilename>bob.pdf</itor:originalfilename>
              <itor:contentbytes encoding="base64"
            mimetype="application/pdf">JVBERi0xLjIgDSXi48/TDQog...
            cmVmDTM1NjU0DSUlRU9GDQ==</itor:contentbytes>
        </itor:itor>
      </data>
    </object>
  </GetObject>
</OA-x>
```

References

1 The author would like to thank Rutger Kramer (NIWI-KNAW) and Laurents Sesink (DENK-KNAW) for reviewing this paper critically.
2 For more information about the DARE programme, see http://www.surf.nl/themas/index2.php?oid=18.
3 For more information about NIWI, see http://www.niwi.knaw.nl.
4 For more information about KNAW, see http://www.knaw.nl.
5 DAREnet has been active since January 2004. On this website, http://www.darenet.nl, it is possible to search the various institutional repositories.
6 For more information about i-Tor, see http://www.i-tor.org. The i-Tor source code is available on SourceForge, at http://sourceforge.net/projects/i-tor.
7 For information about the OAI-PMH protocol, see http://www.openarchives.org/OAI/openarchivesprotocol.html.
8 The i-Tor web technology is made up of various open source components and is based upon open standards (OAI). Built in Java, it is platform-independent and uses a MySQL database. i-Tor is Linux-based and contains software developed by the Apache Software Foundation. It can be implemented as an independent package. Between now and

January 2005 the i-Tor software is being converted to a modular plug-in architecture. This is being done using the Eclipse framework originally developed by IBM. For more information, see http://www.eclipse.com
9 See the i-Tor website, http://www.i-tor.org, for a list of those sites built using i-Tor.
10 For more information about the EVA project, see http://www.eva-eu.org.
11 The Alfred Wegener Institute for Polar and Marine Research is a member of the Helmholtz Association of German Research Centres.
For more information, see http://www.awi-bremerhaven.de.
12 For more information about E-laborate, see http://www.e-laborate.nl.
13 The home page of the electronic publication Studies in Mycology is http://www.studiesinmycology.org.
14 The DARE project Dissertations Online, a joint initiative by the University of Twente, Erasmus University Rotterdam and the University of Groningen, has its
own i-Tor website, http://www.scripties-online.nl.
15 For more information about this protocol, see http://www.openarchives.org/OAI/openarchivesprotocol.html.
16 For more information about OA-x, DIDL and METS, see http://www.i-tor.org/oa_x/retrieving_objects/.
17 For more information about MPEG 21, see http://www.chiariglione.org/mpeg/standards/mpeg-21/mpeg-21.htm. There are several internet sites on DIDL.
18 For more information about METS, see http://www.loc.gov/standards/mets/.

Other Resources

- Arjan Hogenaar. Eindrapportage Project OA-x ("Project OA-x End Report"). NIWI-KNAW, 21/04/2004.
- Marc Evers, i-Tor naar een modulaire plugin-architectuur met Eclipse ("i-Tor: towards a modular plug-in architecture with Eclipse"). In the electronic newsletter Reposi-Tor, no. 5, September 2004.
- Rutger Kramer; MPEG-21 DIDL Approach; OA-x Approach (http://www.i-tor.org/oa_x/retrieving_objects). July 2004.

https://doi.org/10.26069/greynet-2026-000.613-tgj

Building an autonomous citation index for GL: RePEc, the Economics working papers case *

José Manuel Barrueco (Spain)
Thomas Krichel (United States)

Abstract
This paper describes an autonomous citation index named CitEc that has been developed by the authors. The system has been tested using a particular type of grey literature (GL) i.e. working papers available in the RePEc - Research Papers in Economics - digital library. Both its architecture and performance are analysed in order to determine if the system has the quality required to be used for information retrieval and for the extraction of bibliometric indicators.

Introduction

The main characteristic that differentiates the scientific literature from other literary representations is the relationship between documents established trough citations and bibliographic references. The scholarly work can't exist on its own. It must always be related to documents in the same subject area that have been published earlier on. In this way we can see the literary corpus as a complex semantic network. In that network, the vertices are documents and the edges are citations and references.

It is important to differentiate between citations and references. Citations are referrals that a scientific work receives from other documents published later on. References are referrals that one document makes to other works published before.

In the 1960s Eugene Garfield developed the first tool devoted to the representation of relationships between scientific documents: the Science Citation Index. Since then, citation indexes have become an important study tools in some areas. In Scientometrics, citation indexes have become an essential tool for the evaluation of scientific activity. In Information Science researchers have studied the possibility of browsing the scientific literature using references and citations. In this way, once an interesting document has been found, it would be possible to use its references to find similar ones.

Compiling large-scale citation indexes for printed literature, using human labor, has been an expensive task. In the past only the ISI (Institute for Scientific Information) has carried out this type of work. However, nowadays all scientific documents are generated in electronic form. If they are available on the Internet this allows the possibility of extracting the references automatically. The references of a scientific paper identify the cited documents and create the appropriate links if they are available in electronic format. With such system the costs would be dramatically reduced and new indexes covering new document types (i. e. grey literature) could arise.

The pioneers in this research area were Steven Lawrence and C. Lee Giles with the CiteSeer autonomous citation index (ACI) for Computer Science. They define an ACI as a system which ''can automatically create a citation index from literature in electronic format. Such a system can autonomously locate articles, extract citations, identify citations to the same article that occur in different formats, and identify the context of citations in the body of articles. The viability of ACI depends on the ability to perform these functions accurately''. In this paper we describe a similar system called Citations in Economics (CitEc). This system uses CiteSeer technology to automatically build a citation index for documents contained in the RePEc (Research Papers in Economics) digital library.

The remainder of this paper is organized as follows. Section two describes the RePEc data set, which has been used as test bed for the citation index that we have developed. Section three describes the CitEc architecture. Section four is devoted to

* First published in the GL6 Conference Proceedings, January 2005

the analysis of the system performance in order to determine whether it could be used to extract bibliometric indicators. Otherwise it would be limited to information retrieval. Section five concludes the paper.

RePEc: a digital library for Economics

RePEc (Research Papers in Economics) is the largest decentralized non-commercial academic digital library in the world. Its home page is http://repec.org. A gateway that is compatible with the OAI-PMH (Open Archives Initiative, Protocol for Metadata Harvesting) (http://www.openarchives.org), which is available at http://oai.repec.openlib.org. RePEc describes two types of documents: grey literature, namely working papers, and articles published in peer-reviewed journals. In November 2004 there were 140,000 working papers and 144,000 articles. RePEc is based on a distributed architecture where research institutions worldwide share information about the documents they publish. The details are contained in two documents: the Guilford Protocol and ReDIF (Research Documents Information Format).

The Guilford Protocol (GP), named after the town where it was created, is a set of technical requirements that an institution should accomplish to become a member of the RePEc community. It covers only institutional collaboration, i.e. individuals cannot join RePEc. There are two ways for an institution to participate in RePEc: archives or services. Archives collaborate by providing bibliographic information about the documents their institution publishes. They also may provide the full-text of these papers. Technically an archive is a space in the hard disk reachable by an HTTP or FTP server. There, files containing bibliographic information are stored. The structure of this space is defined in the GP.

The second pillar of RePEc is ReDIF. All data transmitted between archives and services is encoded in a specific bibliographic format named ReDIF. ReDIF was created to meet the RePEc needs and therefore it is not aimed to become a widely used format for interchange of bibliographic data between libraries. Its main characteristic is that it is simple enough to be used by non-technical people outside of the library world. This is because administrative staff or academics without knowledge of library procedures maintain most archives.

ReDIF allows one to describe working papers, articles in journals, physical persons and software components. A template made up of several fields like a traditional database record describes each object. They use an attribute: value syntax. Fields can be optional or mandatory. The main mandatory field is the Handle, which holds a code that identifies the object within the RePEc dataset.

RePEc was created in 1997 from a collaboration of several projects working on electronic document dissemination in the discipline. Since then the number of institutions collaborating as archives has been increasing. At the time of writing, in November 2004, there are 413 archives. The number of documents by archive depends on the kind of institution it belongs to. For example, the prestigious NBER (National Bureau of Economic Research, USA) provides a very large archive describing all the 10,897 papers it has published.

Metadata as it appears in the archives is of little utility for researchers. Further processing is needed to take the information and present it in a user-friendly way. This is the objective of the user services. User services are the main way a user works with RePEc. A complete list of user services can be found at the RePEc home page. They add value to the data provided by the archives in several ways:

- Scanning for new data and creating an awareness system to announce new additions.
- Creating a searchable database of papers.
- Creating some type of filtering service to help the user in the selection of the most relevant documents.
- Building a citation index allows to create an additional user service that has citations as its prime focus. This is done in the CitEc project. Its home page can be found at: http://netec.ier.hit-ac.jp/CitEc.

Citations in Economics architecture

The CitEc architecture is based in two main elements as it is shown in Figure 1. First, we have a knowledge base1 where all authoritative metadata about RePEc documents is stored. This base represents the main improvement we have implemented in the CiteSeer software. The quality of the bibliographic references provided in the papers is variable. For instance, it is usual to find different forms for the same author names, journal titles, etc. We use the knowledge base to complete and improve the quality of the references with metadata provided by the publishing institutions. Secondly, we have a series of three software modules, one for each step in the reference linking process:

- Collecting metadata and documents' full text.
- Parsing documents in order to find the references section, to identify each reference, and to extract their elements (authors, title, etc.).
- Linking of references with the document full text they represent (if available on RePEc2).

It is important to note that each module is based on the output of the previous one. In this way, the successful processing of each document implies to successfully surpass the sequence of three levels.

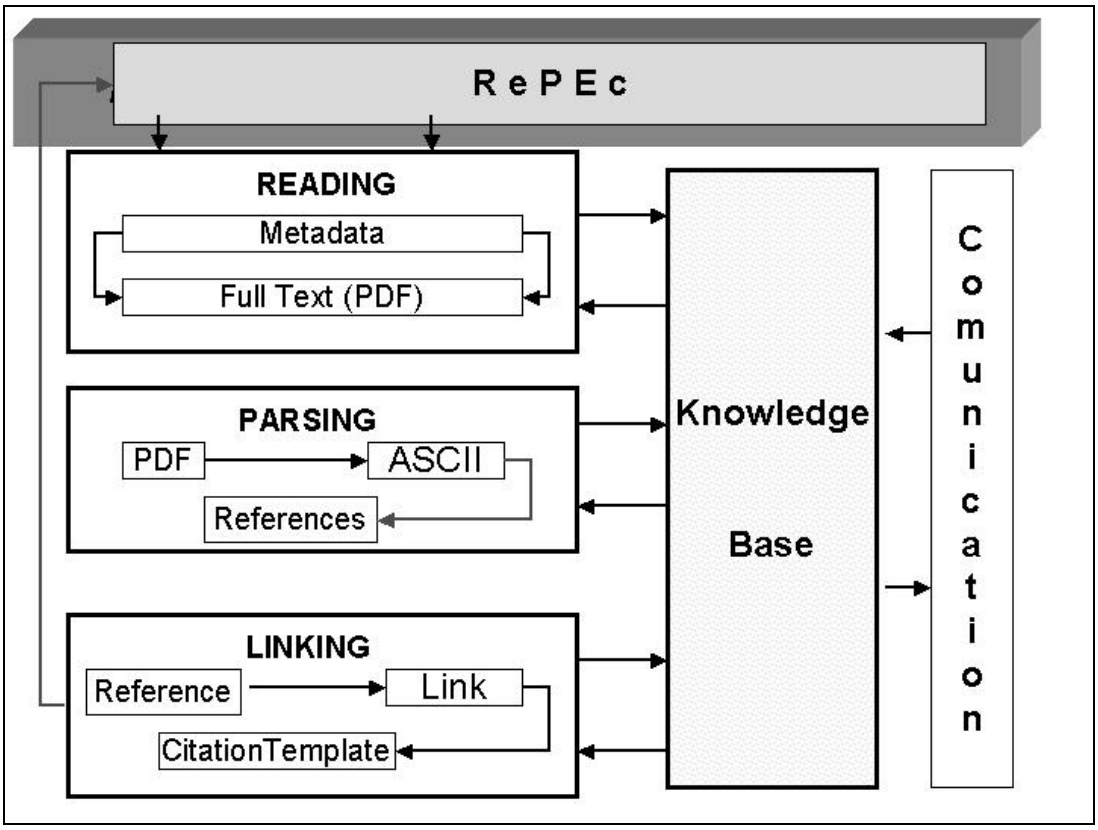

FIGURE 1: CitEc: Architecture

Collecting

Collecting involves three different steps: (1) to collect the documents' metadata, (2) to download the documents' full text and (3) to convert them to a 'parseable' form.

The metadata quality varies from archives. There are archives that provide very complete records for each paper, including abstracts, JEL (Journal of Economics Literature Classification) codes, etc. On the other hand, other archives may only provide titles and authors. There are three main problems with the metadata that seriously affects the processing of papers:

- The absence of publication dates. This field is optional in ReDIF and some archive maintainers don't use it. In our research this data is fundamental because the publication year is one of the attributes we use to check whether a citation goes to

a RePEc document or not. Fortunately the publication year usually forms part of the working paper number. Most series are numbered like: 9901, 9902 ... Taking advantage of this convention, we have developed procedures that guess the year from the paper numbers.

- The format in which the author names are written. ReDIF requires that each name be placed in a different field but some archive maintainers write all authors in a single field, separated by some punctuation. We have also developed procedures to cope with such problems and to correct them as far as possible.
- Wrong URLs. The URLs provided by the archives to retrieve the documents full text are incorrect. This is a rare but serious problem. If we cannot access the paper is not possible to circumvent the problem as we have done with the other cases.

We are working with a distributed library of metadata. There is no a single place where all full text documents live. They are dispersed in multiple servers from multiple institutions. Therefore the second step is to download to our hard disk those documents that are available in full text. This is done by going trough each archive, reading the bibliographic information and, if a File-URL field is found, retrieving the resource contained in the URL. Usually such resource will be the document itself, but in some cases archive maintainers could point the URLs to abstract pages. In these cases the paper will be discarded.

Once the document is saved in our hard disk, we start the conversion process. First, we check if the full text file is compressed. If that is the case, a decompression algorithm is used. Second, we check the file format. Only PDF and PostScript documents are accepted at the moment. Fortunately both are quite popular formats in Economics. More than 95% of the RePEc documents are in either PostScript or PDF.

The last step is to convert the document from PDF or PostScript to ASCII. For this purpose, we use the software pstotext developed by Andrew Birrell and Paul McJones as part of the Virtual Paper project3.

Parsing

Parsing is the most complicated process. Authors usually construct references in a variety of formats, even within the same paper. In addition disciplines vary with respect to the traditions in the way citations are marked in the documents.

Due to the importance of the parsing process we decided to start with software that has been already tested rather than develop new software from scratch. Our choice has been CiteSeer by S. Lawrence, Kurt Bollacker and C. Lee Giles.

CiteSeer is able to identify the part of the document containing the list of references. Then it can split the list into different references. Finally it parses each reference to find the elements. At the moment it only identifies the publication year, the title and the authors. However, as we will see, these four elements are enough for our purposes.

Linking

Once we have parsed the documents, the next stage is to look if some of the references successfully found go to documents identified in RePEc. In such cases, some type of link between both documents should be established. We are doing that by comparing each reference successfully parsed, with the authoritative metadata stored in the CitEc knowledge base. At the moment we consider that a reference represents a RePEc document when:

- The parsed reference title and the title in our metadata collection are close enough.
- The publication year of both items is the same.

In this process we take each reference, extract the parsed title and convert it to a normalized version called key title. Here all multiple spaces and articles are removed and upper case letters are converted to lower case. Then we select from our knowledge base of metadata all documents that contain in their title all the words of the reference key title. All selected papers are suspect of being the cited document. In a second step we compute the Levenshtein distance of each suspect title with the reference title. If this distance is greater than 8% of the title length, the suspect

document is rejected. Finally, we check if the publication year of the suspect papers and the reference is the same. If this is the case we assume that the reference is to the document we have.

Internal Evaluation

In this section we provide a detailed description of errors detected in the processing of RePEc documents in order to determine if our autonomous citation index could be used to provide bibliometric indicators or to assist in information retrieval.

In order to evaluate the system behavior we define a series of stages in the reference extraction and linking process that every paper should pass. In this way, the initial stage for all documents is ''not processed''. It will be changed to the final stage of '' linked'' for papers, which have successful passed all stages in the reference linking process. If the process fails, an error status describing the problem detected is associated with the document. All information about document status and errors is recorded in the knowledge base.

The current version of the system is dated August 2004. It contains 175,452 metadata records with information about electronic documents. Such documents are distributed in the 1591 series coming from 378 institutions worldwide contribute to RePEc. The number of documents per institution ranges from those that only provide one or two documents to those with a national scope, which provide documents coming from several institutions.

We face three problems when downloading documents. Firstly, there are institutions that charge for access to the full text of their publications. In such cases the documents are simply ruled out. We found 51,418 documents with restricted access. That cut down the number of documents to be processed to 124,034. The lion share of the restricted documents comes from commercial publishers and the JSTOR project. Secondly, we found in the metadata wrong URLs to the documents' full text. That means the documents are not found at the specified location due to an error in the metadata provided by the archives. Finally, with the error ''bad document'' there are a variety of problems. For instance, documents digitalized as images, even using the PDF format, and URLs that instead of pointing to the document's full text go to an abstract page. This practice is not allowed in RePEc but some institutions work in this way to make sure their web sites get as much hits as possible. Table 1 shows the ratio of not available documents.

TABLE ONE – Documents' availability

Error	*Documents*	
Restricted	51418	28%
Not found	4221	2%
Bad document	7702	4%
Available Documents	121111	66%

Once the 121,111 available documents have been downloaded the system starts the conversion from the original PDF or PostScript formats to ASCII. In this processing step we found three possible problems: ''incompatible format'' when the format of the file containing the paper is not PDF or PS, ''conversion error'' when the program that convert the formats fails or ''no references'' when even having a text version of the paper, the system has been unable to find a references section in the corpus of the document.

Of the 112,111 documents 61,474 have been successfully converted to text format. Table 2 shows the error distribution for this particular stage.

TABLE TWO – Errors found in the conversion process

Error	*Documents*	
Conversion error	10304	9%
No English	2062	2%
No references	24663	22%
Incompatible format	13708	12%
Converted Documents	61474	55%

At this step it is important to note the large number of documents in which the process of conversion has failed. An initial conclusion to be taked into account in future system updates would be the need of testing new conversion programs.

61,474 documents were parsed in order to locate and identify their bibliographic references. Documents in which the number of references identified by the system is greater than seventy are discarded. In such cases is quite probably the process has failed since such a large bibliography is unusual. As it is shown in table 3 almost the 90% of documents stay below the limits.

In total 1,165,075 references have been identified in 53,201 documents. That represents an average of 22 references by document.

The linking module is the last one in the process. It is in charge of creating a link between each reference correctly parsed and the full text of the document it represents if such document is available in RePEc. 307,094 out of the 1,165,075 references identified are representation of RePEc documents.

In conclusion, 44% of documents available in RePEc were successfully processed. That is, the system was able to extract and link their references. More than half of the documents could not been linked for different reasons. The most important cause of problems is the conversion from PDF to text formats. The second most important is that the system was unable to find the references list in the 12% of the documents. Since it is unusual to find a scientific document without bibliography, we could conclude that the algorithm of analysis needs to be considerable improved in order to extract bibliometric indicator with enough quality to be used in bibliometric studies.

TABLE THREE
Error Frequency

Wrong number of references	2081	13%
Correctly parsed	53224	87%

Conclusion

To sum up, in this paper we have described a system that makes it possible to automatically extract citation data of documents from a distributed digital library. We have designed a procedure to automatically retrieve the documents' full text from the servers and extract the citation data. Whereas this procedure has been proved successful, a few remarks should be taken into account for future work:

- The collaboration of archives maintainers is a key point to allow a correct administration of the system. Good metadata is essential to obtain relevant results in the citation linking process. The main problem we face is wrong URLs to the documents' full text. This will make impossible to analyze the documents. To solve that we are planning to automatically inform the maintainers about each document that could not be downloaded from their archives.
- Better conversion programs from PDF to text are needed. We work with files generated by a wide range of applications. It is possible even to find scanned documents saved as simple images. As a result we need to use a powerful tool that would allow us to obtain a usable text representation of the document.
- The parsing algorithm could be clearly improved. In our example it was able to parse correctly only the 75% of the references. While this can be an acceptable rate of errors when working with reference linking, it is not enough to create more complicated applications like the bibliometric analysis of a discipline.

References

Krichel, T. Guildford protocol.
ftp://netec.mcc.ac.uk/pub/NetEc/RePEc/all/root/docu/guilp.html.

Krichel, T. Redif (research documents information format).
ftp://netec.mcc.ac.uk/pub/NetEc/RePEc/all/root/docu/redif_1.html.

Lawrence, S., K. Bollacker, and C.L. Giles (1999). Indexing and retrieval of scientific literature. – In: Eighth International Conference on Information and Knowledge Management, CIKM99, pp.139-146.

Notes

1. The precise detail of this base is beyond the scope of this paper.
2. Linking of documents out of our dataset, using technologies based in DOI identifiers, in the same way it is done in CrossRef at the moment, is in our to-do list.
3. http://www.research.compaq.com/SRC/virtualpaper/home.html

https://doi.org/10.26069/greynet-2026-000.613-tgj

On the News Front

EL NIÑO AND GREY LITERATURE
A CORDIS NEWS INTERVIEW WITH DR. SVEN THATJE

INCO project helps coastal communities to cope with El Niño

El Niño, the name given to the periodic warming of the Pacific Ocean off the western coast of South America, and its associated cold phase (La Niña), both have a significant impact on the marine biodiversity of coastal areas in Chile, Peru and Argentina, as well as the communities that exploit these resources.

Given the importance of local fisheries to the domestic economies of these countries, a large number of studies have already been carried out to try to identify the effects of El Niño on in-shore ecology. However, much of this data can only be found scattered among the so-called 'grey literature', and analysis of the mechanisms that cause the studied effects is also lacking.

That is why, having identified a high degree of overlap in a number of smaller projects being carried out in this area, and following decreases in national funding for such initiatives, the EU has chosen to fund a four year project aimed at integrating the available knowledge on the effects of El Niño and La Niña on coastal marine environments and resources.

The CENSOR project (Climate variability and El Niño Southern Oscillation: implications for natural coastal resources and management) is funded under the international cooperation priority of the Sixth Framework Programme (FP6). It brings together 13 institutes from six countries - Chile, Peru, Argentina, Spain, France and Germany.

As project partner Sven Thatje, from the Alfred Wegener Institute for Polar and Marine Research in Bremerhaven, Germany, explained to CORDIS News: 'CENSOR was driven by scientists and postdocs keen to integrate the various scattered bilateral cooperations into a larger, multidisciplinary approach to coastal management. The project has a high standard of scientific excellence and is also linked to the needs of South American coastal communities.'

According to Dr Thatje, the most important part of the consortium's work relates to the compilation and analysis of existing data, whether they take the form of national scientific papers, local reports, or news articles. He estimates that the grey literature contains data that would take decades to compile from scratch. 'Once it has been analysed and compiled, we aim to make that information available in a single online location so that people can consult and even contribute to the data themselves,' Dr Thatje added.

A thorough analysis of existing knowledge will also alert the team to where the gaps in our understanding of El Niño impacts on marine ecology are. They intend to compare species fluctuation data with information on El Niño/La Niña events in order to begin to understand the underlying mechanisms (temperature changes etc.) that cause such fluctuations. 'If we can do this, we will be able to provide practical advice to local fisheries on the basis of El Niño predictions, for example suggesting that they remove stocks of a particular species while they still can,' said Dr Thatje.

The CENSOR project represents a new approach to understanding the impacts of climate changes such as El Niño, as rather than using oceanographic indicators to make their predictions, the team will use biological indicators such as the presence of invasive warm water species, the death of local indigenous marine species, and reproductive changes. Dr Thatje fully understands, however, that this will not be an easy task: 'We are talking about a complex ecological system subject to natural fluctuations, climate changes and impacts from fisheries, and it is very hard to distinguish the effects of one from the other.' To do so for all species in the four year

lifetime of the project will certainly be impossible, so the team intends to concentrate on species with the greatest socio-economic value to local communities such as scallops.

As well as filling such knowledge gaps and making their findings available to the general public, the team is also working with national policy makers and local fisheries associations to provide practical advice on specific issues, and to suggest mitigation strategies to offset the effects of El Niño/La Niña events. 'These groups are very open-minded and keen for more advice, so CENSOR will certainly fill a need in that regard,' argues Dr Thatje.

CORDIS News asked Dr Thatje why he feels that the EU was keen to fund the CENSOR project, and what positive outcomes it might have from a European perspective. 'The project represents an exchange of capacities between EU and South American scientists. EU fisheries fleets travel worldwide, so any impacts on fisheries in South America will also have an impact on Europe,' he replied.

On a more scientific level, Dr Thatje added, CENSOR aims at improving our understanding of one particular element of climate change, albeit primarily from a South American perspective. Given that climate change is dictated by complex global systems, however, any improvement in our understanding of one element of the process helps to create a clearer picture of the problem as a whole. 'The knowledge we gain in CENSOR will certainly go towards building intellectual capacity in the EU,' Dr Thatje believes.

Ultimately, however, it is the project group itself that Dr Thatje believes is the most significant element of CENSOR. 'Especially with the junior members of the team, we are helping to develop future academics in South America. Being involved in CENSOR, they see that there is a clear connection between their own scientific excellence and the needs of their countries and local communities. These postdocs are the scientific leaders of the future, so it's also good to get them used to national and international cooperation, and it is a very rewarding experience for all involved,' Dr Thatje concluded.

For further information, please consult the following web address: www.censor.name

Category: Programme implementation
Data Source Provider: CORDIS News interview with Dr Sven Thatje
Document Reference: Based on a CORDIS News interview with Dr Sven Thatje
Programme or Service Acronym: FRAMEWORK 6C; FP6-INTEGRATING; MS-FR C; MS-D C; MS-E C; FP6-INCO
Subject Index: Scientific Research; Environmental Protection; Resources of the Sea, Fisheries; Meteorology; Earth Sciences; Coordination, Cooperation

RCN: 23826

CORDIS RTD-NEWS/© European Communities

https://doi.org/10.26069/greynet-2026-000.613-tgj

Open Access to Grey Resources

INIST-CNRS
Nancy, France
5-6 December 2005

SEVENTH INTERNATIONAL CONFERENCE ON GREY LITERATURE

Conference News

Summer 2005

Communication is the key that opens your Business & Organization to new Markets and Meeting Places

GL7 Program Committee meets in Nancy

The GL7 Program Committee meets on June 9, 2005 at the Institute for Scientific and Technical Information (INIST) in Nancy, France. The purpose of this full-day meeting will be to finalize the Conference Program and Schedule. By mid-June, the authors and co-authors will have received notification of their papers' acceptance, the session in which they are scheduled, the time allotted for their presentation, along with further guidelines for submission of their full-text and PowerPoints. The GL Conference Series is based on a Call-for-Papers. This year some thirty-five abstracts from fifteen countries worldwide were submitted for review.

EBSCO, A Welcome Sponsor of GL7

EBSCO Information Services, the world's most prolific aggregator of full text journals is a sponsor of the Seventh International Conference on Grey Literature. EBSCO is a worldwide leader in providing information access and management solutions through print and electronic journal subscription services, research database development and production. EBSCO offers online access to more than 100 databases and thousands of e-journals and e-commerce book procurement. EBSCO has served the library and business communities for more than 60 years. Its International Headquarters in Birmingham, Alabama, USA provides support for all EBSCO operations.

Recipient GreyNet Award 2005

Prof. Keith G Jeffery, CCLRC - Rutherford Appleton Laboratory (United Kingdom) together with Anne Asserson, University of Bergen (Norway) are co-recipients of the GreyNet Award 2005. The award is in recognition for their contribution to the field of grey literature over the past year. Nominations are based on (1) Results from the GL6 Participant Evaluation Forms, (2) Publication of the authors' paper in the GL6 Proceedings, (3) Selection and publication of the authors' paper as a journal article, thus exporting research results originating in the GL Conference Series, (4) Prior history of the authors on the topic of grey literature, and (5) Willingness to attend the GreyNet Award Dinner in order to receive the honor. This year's GreyNet Award Dinner will be held on December 5th 2005 in the Marie-Antoinette Room of the Grand Hotel de la Reine situated on the Place Stanislas in Nancy.

GL7 Hotel Booking and Travel Information

Nancy celebrates this year the 250th Anniversary of Place Stanislas recognized as one of the most beautiful squares in the world. Nancy has been dubbed 'Capital of the Age of Enlightenment' and major exhibitions, shows, and meetings are scheduled throughout the entire year. In order to facilitate travel and lodging to GL7, sheet downloads are now available on the conference website at http://www.textrelease.com/pages/4

ISBNs registered for the GL7 Program Book and Conference Proceedings

- GL7 Program and Abstract Book, Dec. 2005. - ISBN 90-77484-05-1. - (ISSN 1385-2308; No. 7)
- GL7 Conference Proceedings, January 2006. - ISBN 90-77484-06-X. - (ISSN 1386-2316; No. 7)

TextRelease

GL7 Program and Conference Bureau
Beysterveld 251, 1083KE Amsterdam, Netherlands
www.textrelease.com - info@textrelease.com - T/F +31-(0)-20-672.1217

https://doi.org/10.26069/greynet-2026-000.613-tgj

https://doi.org/10.26069/greynet-2026-000.613-tgj

About the Authors

Nikkia Anderson, a 2004 graduate in Computer Science from Bowie State University, is a Computer Programmer with Information International Associates, Inc. (IIa). Her current projects include the development of a core metadata set for the Goddard Library; 508 compliancy, and developing procedures for digitizing a collection of NASA Balloon Technology documents.
Email: nikkia@pop200.gsfc.nasa.gov

José Manuel Barrueco is a PhD student at the Universidad Politecnica de Valencia. He also works as assistant librarian at the Social Science Library of the University of Valencia. Since 1994 he has been involved in different international projects devoted to improve the scholarly communication process using electronic resources. In Economics he is one of the founders of RePEc (Research Papers in Economics). In Economics he has developed CitEc (Citations in Economics) the first autonomous citation index in this discipline. His work on LIS has been centered in DoIS (Documents in Information Science). He is also one of the founders of E-LIS the largest open archive in the discipline.
Email: barrueco@uv.es

Marie-France Claerebout
She holds a DESS post-graduate degree in information and documentation sciences, a European masters in software engineering, and a DEA (equivalent to M.S./M.Sc) in theoretical computer science from the University of Lille. Since 2000, she is a librarian in charge of theses and grey literature at the University Library of Lille, Science and Technology Division and is responsible for the realization of the digital library, Grisemine. She also coordinates the electronic dissemination of the local theses. She is a member of the AFNOR committee for normalization of metadata for electronic theses and a member of the technical committee (COUPERIN consortium) for the valorization of local grey literature.
Email: Marie-France.Claerebout@univ-lille1.fr

Elly Dijk is a Policy Officer at the Netherlands Institute for Scientific Information Services (NIWI), part of the Royal Netherlands Academy of Arts and Sciences (KNAW). For the past three years she has been involved in the development of the open source web technology i-Tor, Tools and Technology for Open Repositories and leads various i-Tor projects. She is a member of the DARE & Co (Communication) working group of the Digital Academic REpositories (DARE) programme. Before taking up her current post, Dijk was Senior Scientific Information Specialist in the Research Information Department at NIWI. She is a graduate of the University of Amsterdam, where she studied Human Geography and Documentary Information Sciences.
Email: Elly.Dijk@niwi.knaw.nl

Julia Gelfand has been a librarian with the University of California, Irvine Libraries since 1981. She has been tracking the grey literature movement since the late 1980s and has participated in all of the previous GL conferences and has published and presented widely on different topics in grey literature. Her particular interests are in scholarly communications, electronic publishing, collection development, bibliography of science and technology, and she thinks that with more emphasis on networking and digital libraries, Grey Literature has a very interesting future. Gelfand is currently the chair of the IFLA Science & Technology Section and vice-chair/chair-elect of the ALA ACRL Science & Technology Section. Email: jgelfand@uci.edu

Andrea C Japzon earned a Master of Library Science from Florida State University in 1994 and in the same year began working as a librarian at The New York Public Library. In 1999, she joined the library faculty at Hunter College of the City University of New York. While at Hunter, she completed an MA in geography. As of 2004, she has worked for the NASA Goddard Space Flight Center Library contractor, Information International Associates, Inc. She is a member of the Society of Woman Geographers and the American Library Association. Email: ajapzon@pop200.gsfc.nasa.gov

Thomas Krichel, born in Saarland in 1965. Studied Economics and Social Sciences at the universities of Toulouse, Paris, Exeter and Leicester. Between 1993 and 2001 he lectured in the Department of Economics at the University of Surrey. In 1993 he founded NetEc, a consortium of Internet projects for academic economists. In 1997, he founded the RePEc dataset to document Economics. In 2000, he held a visiting professorship at Hitotsubashi University. Since January 2001, he is assistant professor at the Palmer School of Library and Information Science at Long Island University. In 2002, he founded the rclis dataset, a clone of RePEc for Computing and Library and Information Science and the Open Library Society.
Email: krichel@openlib.org

LeRoy LaFleur is a Social Sciences Bibliographer at Cornell University's Albert R. Mann Library where he is responsible for developing the library's collection of social and life sciences materials and serves as liaison to a variety of social sciences departmental units. Additionally, he provides reference and consultation services to Cornell students, faculty and staff and teaches course related and application based workshops. He holds an undergraduate degree in Sociology from Michigan State University and a Masters of Library and Information Science from the University of Wisconsin-Madison. Email: ljl26@cornell.edu

Nathan Rupp is a metadata librarian at Cornell University's Albert R. Mann Library. In addition, he assists Cornell students, faculty and staff with research through the library's reference program and teaches workshops on various computer applications. He is currently researching how webblogs fit into library collections, the usefulness of the traditional library online catalog as a data repository for supporting other library projects, and how a repository for storing metadata schemes and transformations might help librarians reuse metadata tools to more efficiently develop digital library projects. He has an undergraduate degree in history and graduate degree in library science from San Jose State University in California. Before coming to Cornell, he worked for academic libraries in Indiana and Pennsylvania. Email: nar25@cornell.edu

https://doi.org/10.26069/greynet-2026-000.613-tgj

Notes for Contributors

Non-Exclusive Copyright Agreement

- I/We (the Author/s) hereby provide TextRelease (the Publisher) non-exclusive copyright in print, digital, and electronic formats of the manuscript. In so doing,
- I/We allow TextRelease to act on my/our behalf to publish and distribute said work in whole or part provided all republications bear notice of its initial publication.
- I/We hereby state that this manuscript, including any tables, diagrams, or photographs does not infringe existing copyright agreements; and, thus indemnifies TextRelease against any such breach.
- I/We confer this copyright without monetary compensation and with the understanding that TextRelease acts on behalf of the author/s.

Submission Requirements

Manuscripts should not exceed 15 double-spaced typed pages. The size of the page can be either A-4 or 8½x11 inches. Allow 4cm or 1½ inch from the top of each page. Provide the title, author(s) and affiliation(s) followed by your abstract, suggested keywords, and a brief biographical note.
A printout or PDF of the full text of your manuscript should be forwarded to the office of TextRelease. A corresponding MS Word file should either accompany the printed copy or be sent as an attachment by email. Both text and graphics are required in black and white.

REFERENCE GUIDELINE

General

i. All manuscripts should contain references
ii. Standardization should be maintained among the references provided
iii. The more complete and accurate a reference, the more guarantee of an article's content and subsequent review.

Specific

iv. Endnotes are preferred and should be numbered
v. Hyperlinks need the accompanying name of resource and date; a simple URL is not acceptable
vi. If the citation is to a corporate author, the acronym takes precedence
vii. If the document type is known, it should be stated at the close of a citation.
viii. If a citation is revised and refers to an edited and/or abridged work, the original source should also be mentioned.

Example

Youngen, G.W. (1998), Citation patterns to traditional and electronic preprints in the published literature. - In: College & Research Libraries, 59 (5) Sep 1998, pp. 448-456. ISSN 0010-0870

Review Process

The Journal Editor first reviews each manuscript submitted. If the content is suited for publication and the submission requirements and guidelines complete, then the manuscript is sent to one or more Associate Editors for further review and comment. If the manuscript was previously published and there is no copyright infringement, then the Journal Editor could direct the manuscript straight away to the Technical Editor.

Journal Publication and Article Deposit

Once the journal article has completed the review process, it is scheduled for publication in The Grey Journal. If the Author indicated on the signed Copyright Form that a preprint of the article be made available in GreyNet's Archive, then browsing and document delivery are immediately provided. Otherwise, this functionality is only available after the article's formal publication in the journal.

https://doi.org/10.26069/greynet-2026-000.613-tgj
TGJ Volume 1, Number 2 2005

An International Journal on Grey Literature

'REPOSITORIES – HOME2GREY'

SUMMER 2005 - VOLUME 1, NUMBER 2

Contents