



*Greater visibility and application of research through global networks of
Open Access repositories*

Sustainable Best Practices for Populating Repositories

Preliminary Report

Produced by the Confederation of Open Access Repositories (COAR)

May 2012

<http://coar-repositories.org>

Introduction

In early 2012, the Confederation of Open Access Repositories (COAR) began a project to collect and disseminate sustainable, replicable best practices related to populating repositories.

The focus of this project is on long-term, sustainable activities, strategies, and operations – i.e. ways in which organizations can improve deposit rates or increase the content in their repositories on a permanent basis.

- Lessons learned: what's worked and what hasn't?
- Workflows or strategies to gather content and populate repositories over time.
- Long-term staffing, operational plans, or campus partnerships that have led to consistently higher rates of deposit.
- Automated processes to gather content and populate repositories.
- Any other long-term activities or operations that resulted in higher levels of article deposit.

Outcomes

Case studies will be posted on the COAR website (<http://www.coar-repositories.org>) and COAR will continue to collect and share these types of best practices for content recruitment on an ongoing basis. A report summarizing the sustainable practices, analysis and conclusions will also be published.

Some preliminary conclusions

- Populating repositories remains a challenge and as a result many institutions are using a multi-faceted strategy to actively recruit content.
- Most content recruitment practices are fairly resource intensive and involve staff resources.
- OA mandates do not result in high deposit rates without support for depositing content.
- Although the focus of this project is on recruiting journal articles, it is apparent that many repositories are expanding their scope to collect a much wider range of content including data and digitized collections.
- So far, there are no “magic bullets” for fast and easy populating of repositories.

Summary of sustainable practices for populating repositories

1. Researcher advocacy
2. Using usage statistics to increase deposits
3. Rights checking and content submission services
4. Automated downloading of citation data (from citation databases or institutional publication lists)
5. Full text harvesting (from institutional web sites or disciplinary repositories)
6. Repositories and Research Assessment Exercises (REAs)
7. Direct deposit by publishers

Case studies

I. Researcher advocacy

I.1. hita-hita: Institutional OA Advocacy in Japan

The term “hita-hita” was employed to signify something that penetrates gently and smoothly. As of 2011, a total of 760,000 research papers are accessible through IRs across Japan. Open Access (OA) is not politically forced at state level or institution level. Instead, we have built repository manager-level lateral solidarity to share experience, exchange various creative ideas, add improvements and disseminate our ideas, thereby highly developing respective IRs.

The hita-hita report introduces large and small ideas that have enhanced penetration of the OA movement and developed IRs in Japan, and it also outlines the activities of DRF, which has supported such ideas. A number of Japanese institutions are contacting staff directly to introduce and discuss the benefits of Open Access and depositing their papers into a repository.

More information:

<http://drf.lib.hokudai.ac.jp/drf/index.php?plugin=attach&refer=Digital%20Repository%20Federation%20%28in%20English%29&openfile=hitahita2011.pdf>

Targeting scholars directly at Institutional Repository der Universität Konstanz

The more directly repositories target scholars, the greater the prospects of success. Authors who have already made work openly accessible are often not only prepared to continue to do so but are also frequently willing to recruit content on the repository's behalf. For repository operators it is particularly advantageous to win the support of leading researchers who are willing to act as multipliers in this way.

Another strategy is to target those scholars who make a lot of their work openly accessible on their own personal website anyway. It is safe to assume that they will be open to depositing their work on a document server. This type of contact can be used to recruit further content. Furthermore, it is important to take advantage of good opportunities whenever they arise. For example, an enquiry from a researcher from the University of Konstanz about a copyright issue relating to the redesign of his website on which he intended to make full-text versions of his work openly accessible, led to the organisation of an information session for several branches of the faculty on the subject of Open Access and copyright. This resulted in the recruitment of over 500 publications from faculty members.

More information:

<http://open-access.net/chen/usefulinformationfor/repositoryoperators/einwerbenvondokumenten/>

2. Using usage statistics to encourage deposits

2.1. Research@StAndrews:FullText

We are delighted that we are now able to show usage statistics for content in our institutional repository, Research@StAndrews:FullText. Thanks to the services provided by the Scottish Digital Library Consortium (SDLC), we now have 'page views' and 'downloads' from Google Analytics integrated into our DSpace platform and publicly available.

Visitors to the repository can see statistics for individual items. Statistics initially show how many times the metadata for an item has been viewed, and from the 'View Statistics' button users can see a breakdown by date range, plus the number of times the full text has been downloaded.

More information:

<http://univstandrews-oaresearch.blogspot.com/2012/02/new-public-stats-now-available-for.html>

3. Rights checking and submission services

3.1. Institutional Repositories Supporting Cyberscholarship: Building Content, Griffith University, Queensland, Australia

The critical success factors for Griffith Research Online have been:

- "Harvest", i.e. find content published by Griffith authors and then ask authors for relevant files
- Allocate one staff member to contact publishers' for permission as well as keep abreast of which publishers now allow publisher PDF version
- Upload files on behalf of Griffith authors
- Undertake all copyright checking
- Provide easy access to a range of statistics relating to each publication by a given author
- Demonstrate how searching in Google Scholar returns an entry in GRO
- Utilise both marketing and support strategies which are tailored to meet the needs of different "cultures", e.g. disciplines

More information:

<http://ausweb.scu.edu.au/aw08/papers/refereed/richardson/paper.html>

3.2. The GALILEO Knowledge Repository

Eleven institutional partners in the state of Georgia are working together to build a state wide institutional repository (IR) called the GALILEO Knowledge Repository.

Three services are under development to gather USG faculty's scholarly works:

- Rights assistance: Through the Scholarly Communication and Digital Services Department, the GT Library provides rights management advice to members of the GT community. They will extend their advisory services to the other GKR campuses via the partner library's liaison. The service will be a contributed cost match of staff time.
- Digitization: A multitude of scholarly works are locked up in print, which makes it difficult for GKR partner sites to disseminate them digitally. Scholarly materials like theses and dissertations, research reports, working papers, and research data are not easily accessed in print. Student assistants (IMLS funded) and existing DLG staff will carry out digitization. The GKR Content and Metadata Committee will conduct an approval process for digitization requests, as well as determine priorities for digitization efforts.

- Content submission: Each GKR partner site has final responsibility for collecting, submitting, and organizing their IR content; however, many do not have the appropriate staff or training to do this work. The GKR project team has created and approved the *Metadata Guidelines for the Galileo Knowledge Repository*, a DSpace enabled, modified version of Dublin Core. The GKR Content and Metadata Committee maintains the *Guidelines*, reviews all content submission and metadata issues, and with the GKR Manager, will provide guidance to the Content Submission Service. Funds from IMLS are budgeted at \$10,000 per year for the GT and UGA libraries to provide services to the other GKR sites needing assistance. Services include interpreting and applying the *Guidelines*, resolving submission issues, and submitting content and metadata. The GT Library will manage the funds and disperse them at the rate of \$20/hour to other GKR partner libraries providing metadata and content submission services. These library staff also will interact with the Valdosta State University (VSU) Master of Library and Information Science program, which will provide 2-4 student interns to serve as content submitters for GKR sites.

More information:

<http://www.library.gatech.edu/gkr/>

3.3. EconStor: A RePEc Archive for Research from Germany by Jan Weiland

EconStor is a subject-based repository for economics and business administration maintained by the German National Library of Economics / Leibniz Information Centre for Economics (ZBW). It provides free access to all kinds of scholarly publications, including working and discussion papers, conference papers, journal articles, research reports, and dissertations. The main content so far comes from German research institutions and university departments. But acting as a disciplinary repository EconStor, of course, welcomes any research institution worldwide seeking for a reliable storage and publishing infrastructure for its research papers in the field of economics and business administration – especially those institutions without access to a local repository infrastructure.

In order to achieve these goals we decided to make a “Full-Service Offer” to the editors of publications being considered to be published at EconStor, i.e. the EconStor team organizes the full text upload and metadata recording – free of charge, but based on a publication agreement which is required for copyright reasons.

The idea of building up a ‘national RePEc input service’ was convincing for the German Research Foundation (DFG), that decided to supply some extra funding for the implementation in 2009. The funding enabled us to transfer the RePEc export

interface to DSpace and to prepare additional publications for the integration into EconStor. This includes several 'back files' from the early 1990s, which in some cases had been originally published in formats like Postscript, DVI/TeX, or pure HTML – and are now available in PDF on EconStor and in RePEc.

More information:

<http://www.econstor.eu/>

3.4. Ifremer

82% of Ifremer 2005-2010 publications online for free. These good results are linked to the involvement of Ifremer's library in the operation of this archive. It is the library staff itself which carries out the preparation and registration of publications into Archimer. In addition, in order to enable the recording and dissemination of the greatest number of publications, since we can't count on spontaneous deposit by the authors, we conduct our own monitoring and collections:

- Every week, we harvest the publications written or co-written by Ifremer in the Web of Science®
- Then we study the editor's policy for each of these publications on the Sherpa/Romeo website. If the publisher's policy is not declared, neither on the site Sherpa/Romeo, nor on its own website, we systematically contact the editor to ask permission to record these articles into Archimer
- If the publisher allows the self-archiving of its own PDF files (ex: EDP Sciences, The Company of Biologists), we download the PDF file of the article from the editor's website and then record it into Archimer
- If the publisher restricts the right of self-archiving to the last draft of the publication, we ask the author their own version using automated tools developed by Ifremer. If they can provide us with this version, we create a PDF file from the files sent, which we then register into Archimer
- We contact authors every week to collect their publications. A system recovery, also automated, allows us to remind the authors that we expect their paper for Archimer. This harassment policy causes few calls from scientists fed up with these calls (up to 6 calls over a long period of time), but this work is generally very well received.
- We update all recorded publications with the editor's final version, which visibility is limited to the Ifremer's intranet. In fact, Archimer contains the full text of 100% of articles published by Ifremer since 2000.

More information:

<http://archimer.ifremer.fr/open-access2010.htm>

3.5. Smithsonian Digital Repository

The Smithsonian Research Online consists of two basic components: a list of publications authored by Smithsonian researchers and affiliates, and a corresponding digital repository which contains the actual article or chapter in electronic form. The data which SRO collects is not only used by Institution administrators for research assessment purposes, but is also re-used by webmasters and other Smithsonian offices for reports, presentations and other public information services.

Although there have been over 3000 items submitted in the first year, the vast majority of that content has been ingested to the system by Smithsonian Institution Library staff rather than by authors or creators of the digital objects. Typically authors send their electronic reprints to the SIL on CDs or via some other file transfer. Library staff search for and download corresponding metadata from several available resources including an Institution-wide research bibliography also managed by the library. Metadata and digital object(s) are matched and later uploaded in bulk. This approach imposes somewhat tedious labor on the library staff but it eliminates time that would have been spent on training (and re-training) scholars on the ingest process, etc.

More information:

<http://smithsonianscience.org/2011/02/digital-repository-now-contains-10000-items/>

4. Automated downloading of citation data

4.1. Five Colleges of Ohio Digital Initiatives

Catalina Oyler, Five Colleges of Ohio Digital Initiatives Coordinator, developed, as a part of an Andrew W. Mellon Foundation grant, a procedure for batch loading scholarly article citations (from Web of Science [etc.]/via Refworks) into a DSPACE scholarly article repository. This allowed Oberlin College to efficiently load large numbers of faculty citations for 2010 and 2011 as a means of growing the IR.

More information:

<https://sites.google.com/a/ohiolink.edu/drmc/documentation/batch-submission-from-refworks>

4.2. InterNano repository

It is part of the National Nanomanufacturing Network, which is an alliance of academic, government and industry partners that cooperate to advance nanomanufacturing strength in the US.

One of our strategies to build content for the repository has been to contact these research centers and ask for their lists of publications each year (this is hundreds of articles for each). Each center collects this information anyway for their annual reports to NSF. I have then had a student identify the DOIs and then we just batch import the records into the repository. We have an original taxonomy that I or a student can go through and tag the articles with. We have thought this has been a great strategy to document the field, which is both inter and multi-disciplinary. Our resources serve both academic and industry audiences.

More information:

<http://eprints.internano.org/>

5. Linking repositories with Research Assessment Exercises (RAEs)

5.1. Leveraging research quality assessment exercises to increase repository content - an Australian case study

The legacy of Australia's national research quality assessment process – Excellence in Research for Australia (ERA) - at the University of Wollongong during 2010 was improved integration between the institutional repository and research management systems, and a move towards digitisation of the Higher Education Research Data Collection (HERDC) with consequent flow-on of metadata and digital objects to the institutional repository. Whilst ERA was a diversion from the task of securing open access content through faculty promotion and one-on-one contact with researchers, it nevertheless gave rise to a semi-automated process which promised improved rates of content acquisition.

More information:

<http://ro.uow.edu.au/asdpapers/253/>

6. Full text harvesting

6.1. Harvesting back from subject repositories: The Australian National University, Canberra, Australia

If academics are already making their work available open access elsewhere, your repository can potentially ‘harvest’ this work. This method depends on your university’s technical capabilities and repository content policy. In some instances the repository is supposed to contain a copy of everything the academic has produced. Some repository software will not work unless there is an item attached to the record, so providing a link is not sufficient. Other repositories can point the user to an external source.

There is an argument that if it is available elsewhere why bother also putting it in the repository, but if you want to have a repository that reflects the breadth of work by the academic community (and a repository that looks substantive) then this is an opportunity.

More information:

<http://cairss.caul.edu.au/www/openaccess/dannykingsley/recruitingmaterialforrepositories.pdf>

7. Direct deposit by publisher

7.1. How to increase content in OA Repositories – What can be learnt from the special case of the research project PEER – Publishing and the Ecology of European Research

There are some findings and developments from the PEER project on how to obtain more articles for institutional or subject-based OA repositories, which should be of interest to COAR members. With the following remarks strategies for filling repositories are presented - for the very special case of PEER.

The SWORD5 protocol (Simple Web-service Offering Repository Deposit) deposit mechanism offers a unified ingestion service and guarantees a robust transfer of manuscripts. In PEER, the SWORD protocol is used for article ingestion from the PEER Depot to repositories as it offers the option to transfer content to all partner repositories in a single simultaneous automated transfer process. Not only is this a new application in terms of the transfer of both metadata and full-text articles, it also results in a limited percentage of unknown errors in the transfer process and is thus a useful effort to achieve maximum content.

This procedure for ingestion of material into repositories has applications beyond the project and may, in theory, be adopted by repositories to accept material directly from various sources, e.g. publishers, in the future. In addition to the implementation of the SWORD protocol, PEER has developed a technical infrastructure that has applications beyond the project. This does not only apply for PEER partners but also for other interested parties. An essential ingredient of the PEER infrastructure are the Guidelines for publishers and repository managers on deposit, assisted deposit and self-archiving which were particularly designed to fit the needs of participating publishers and repositories.

More information:

<http://www.peerproject.eu/reports/>

7.2. PubMed Central and UK PubMed Central: Publishers direct deposit into repositories

In July 2008, Nature Publishing Group (NPG) launched the first phase of its Manuscript Deposition Service. The free service helps authors fulfil funder and institutional mandates. As of January 2009, 43 journals published by NPG are included in the service.

Several funding bodies and institutions have introduced mandates that require authors to self-archive articles in publicly accessible archives. While these mandates apply to authors, not publishers, NPG is committed to providing an excellent service to our authors at every stage of their publishing experience. It seemed a natural extension of NPG's self-archiving policy to offer a service to authors to do this on their behalf.

NPG's Manuscript Deposition Service is currently available to authors publishing original research articles in Nature and the Nature research titles and many of the society and academic journals published by NPG. NPG continues to work to make the service available to authors publishing original research articles in the remainder of its society and academic journals.

www.nature.com/authors/author_resources/deposition.html