What can we learn from Europe in our quest for populating our repositories?

Vanessa Proudman, Tilburg University

1. Introduction

This paper seeks to highlight some of the results of a research project entitled *Stimulating the Population of Repositories* commissioned by DRIVER and SURF in 2006 and conducted in 2006/2007.[1][2] The paper was presented on 2 April 2008 at the Open Repositories Conference 2008 in Southampton, UK.[3] Six repository and repository-based service good practices were selected from Europe which showed promising object file growth. Lessons learnt from these cases are to be shared with the international repository community through this paper. The author will highlight a small selection of critical success factors for successfully populating a repository or IR-based service. Reasons for researcher take-up, considering the role of the discipline in successfully populating repositories, and the types of existing services which support the researcher and institution in their work will be explored.

2. Background

2.1 Aim

The aim of the DRIVER research study was to investigate some of the successes in the European repository community as a means to inspire repository managers and policy-makers from different environments and contexts in their quest to better fill their OAI-PMH archives. Various examples of repository or service were under investigation to show the breadth of the challenge and to ensure that most readers of the study can identify with at least one of the models. The models studied were: a departmental repository and its related institutional repository, an institutional repository, a central national archive, an international institutional research repository, a national service based on repository content, and an international subject-based repository service.

2.2 Analytical framework

---

[1] DRIVER - Digital Repository Infrastructure Vision for European Research: http://www.driver-community.eu/

[2] The SURF Foundation, is a funding agency and partnership organisation which serves all Dutch higher education institutions by developing network services and information and communications technology (the JISC or DfG are equivalents). http://www.surf.nl/en/Pages/home.aspx

[3] Open Repositories Conference 2008: http://or08.ecs.soton.ac.uk/

The DRIVER research project *Stimulating the Population of European Repositories* profiled six good practices which demonstrate where the population of digital repositories is gaining ground in Europe. Six repositories and/or services were selected and analysed in six areas: 1) policy issues, 2) organisation, 3) mechanisms and influential factors for populating repositories, 4) services, 5) advocacy and communication and 6) legal issues. These areas are also recurring themes of international and national discourse on the issue of open access and scholarly communication. However, this study goes into more depth into operational issues for the Institutional Repository manager. It does this by highlighting the whys and hows, the critical success factors, the choices made and detailed contexts for lessons learnt to be of real use.

2.3   The selection of case studies

The first milestone in the research was to determine the case studies. Desk research was carried out using the directories OpenDOAR and ROAR to analyse the size of repositories in terms of metadata and full text numbers. Growth patterns and rates were also observed.[45] As a result, a preliminary short-list of European repositories and services was created. Initial telephone interviews were then carried out with those on the preliminary short-list to verify the ROAR and OpenDOAR data. Further questions were posed on growth and take-up by the research community. This resulted in the final selection of six case studies.

Success indicators which were used to determine case study selection were full text numbers, percentage of academic output, striking and/or steady growth data, and take-up by the research community. All cases are, in addition, OAI-PMH repositories and address scholarly output. Cases are neither data archives, nor learning object ones. The selection of case studies was also determined by the objective of highlighting different approaches to aspects such as: population policies, organisational profiles, repository types and services, language content, and geographical distribution.. The examples identified have been chosen to represent different models of repositories and services which have stimulated IR population. Cases are neither typical nor completely unique. For a list of the six cases selected see below. Owing to the scope of the DRIVER Project the study is limited to six cases in the European domain. However, good practices and lessons learnt in the repository field should know no geographical boundaries.

2.4   Interviews

Once cases were selected, interviews were held face-to-face with Library Directors, IR managers and initiators as well as with staff on more operational levels in some cases. Interviews were held in late 2006/early 2007. In-depth case studies were then written targeting IR managers and their policy makers in particular. Interviews were held to investigate the cases' establishment, user relations, organisational set-ups, advocacy programmes, services and other factors which influence repository growth.

---

[4] The Directory of Open Access Repositories – *Open*DOAR: http://www.opendoar.org/
[5] ROAR: http://roar.eprints.org/?action=browse

These case study write-ups are available at
http://www.tilburguniversity.nl/services/lis/driver-population.html

2.5   Publication

This research resulted in the following publication.[6]

> Proudman, V. (2007) The population of repositories In Eds. K. Weenink, L.
> Waaijers and K. van Godtsenhoven, *A DRIVER's Guide to European
> Repositories* (pp.49 - 101) Amsterdam: Amsterdam University Press.
> For more details and full text open access to this chapter, see
> http://dare.uva.nl/aup/nl/record/260224.

3. Six European good practices

The six good practices were selected to represent different models of repositories and services which have demonstrably stimulated repository population.

3.1   Institutional Repositories

3.1.1   University of Minho, Portugal. Minho's university institutional repository has had a broad take-up from its research community. 26 out of the 30 research centres actively contribute to the repository and academic output achieves an 80-90% deposit rate in the areas of biology, bio-engineering and civil engineering. Minho's mandate and financial incentives for research communities have had positive effects on its repository content. In addition, Minho has a developed advocacy programme and support infrastructure which provides help to de-centralised research centre repository communities across the campus.
https://repositorium.sdum.uminho.pt/?locale=en

3.1.2   ePrints Soton and ECS Eprints Repository, UK. A campus-wide institutional repository run by a Library (ePrints Soton) liaises with a university school repository run by a research department (ECS Eprints Repository from the School of Electronics and Computer Science). They both have a common mission but have different methods and challenges, for example with organisation and quality control. The ePrints Soton repository's function as the main source for the university's submission to the UK's Research Assessment Exercise (RAE) is also of interest.[7] The ECS school archive was among the first OAI repositories to be established. Various services are on offer by both.
ePrints Soton: http://eprints.soton.ac.uk/ ECS ePrints: http://eprints.ecs.soton.ac.uk/.
Several are described below in section 5.

---

[6] Excerpts from this publication have been used for parts of this paper.
[7] RAE 2008: http://www.rae.ac.uk/

3.2   National services

3.2.1   Hyper Article on Line (HAL), France, is a central archive repository which brings together French national research results (HAL). This central organisational model's approach and its results are highlighted in this study. It is multi-disciplinary in scope whereas most other known central repositories usually have one specific disciplinary focus. HAL has developed a number of services to face the challenge of supporting the whole French research community.
http://hal.archives-ouvertes.fr/index.php?langue=en

3.2.2   Cream of Science, the Netherlands, is a repository-based service which brings a new quality stamp to IR content. This service aggregates leading national researchers and their output from a number of repositories showcasing national research results. A contributing factor to the success of the Cream of Science has been the creation of four active national people networks: of library directors, repository managers, technical developers, and communication experts.
http://www.creamofscience.org

3.3   International subject-specific service

CERN Document Server is an international research organisation institutional repository (CERN). It adds to its institutional content by harvesting from circa 100 world-wide information sources to create a large information resource for particle-physics. Critical mass has been gained where CERN achieves close to the 80% mark of its academic output on an annual basis. Challenges and opportunities in depositing and reclaiming particle physics content are shared in the study. http://cdsweb.cern.ch/

Connecting Africa, the Netherlands, is a subject-specific service model built on Institutional Repository content. It serves a specific international community of researchers by providing a portal which pulls information together from an array of repositories to enhance networking and access to research.
http://www.connecting-africa.net/

4. Reasons for take-up

It is clear that information about what motivates researchers to contribute to repositories or not, is helpful to determine researchers' needs. The study's six cases were asked why their researchers contributed object files to their repositories and services. A number of researchers supported the general principle of Open Access. However, increasing visibility and impact was the most popular reason for collaborating in repository activities. Otherwise, those who had mandates to abide by stated depositing for reasons of obligation. In the UK case study (University of Southampton) the Research Assessment Exercise (RAE), was mentioned as a driver: the evaluation of departmental output and repository activities were combined into one.

The general preservation function of the archive is also a motivation for researchers to deposit copies of their work. Archives promise to address short to long term preservation issues by: storing copies of data; storing format information as HAL does to prepare for long-term preservation; or as is the case for Cream of Science, where all material deposited in collaborating IRs is archived by the National Library of the Netherlands in the long term.[89] In the case of services which addressed a larger group than one specific institution, researchers also co-operated due to existing professional relations with the service provider as in the case of Connecting Africa for example.

In the case that deposit was undertaken on behalf of researchers by the Library, some reported that their researchers contributed as it meant little effort for them. Last but not least, services such as automated publication lists, RSS feeds, and the provision of new material were mentioned as reasons for depositing material with repositories by various cases. For more information on this, see the next section.

5. Services

Mandatory deposit policies are clearly important for better guaranteeing repository growth. Evidence from this research project has shown that the three institutional repositories, Minho, Southampton and CERN, all have mandatory deposit policies in place. However, all are achieving 40-65% of academic output through direct deposit. All cases admit that 'carrots' are needed as well as 'sticks'. Therefore incentives such as services which answer researcher and institutional needs and problems need to be provided.

Twenty-seven services were specifically mentioned by the six good practices. Seven services were common to three or more cases studied. Seventeen services were unique to one or other of those surveyed. The most popular service mentioned by all was repository search and browse facilities. Five out of six recorded disseminating repository content to other information services be they 1) national ones such as Intute in the UK, 2) broader international ones such as Google and Google Scholar or 3) disciplinary ones such as arXiv.org or RePEc.[10][11][12][13] Five out of six also reported generating automated publication lists from their archives with links to underlying object files. Three or more stated providing bibliographic lists for export to be used on personal web pages or departmental websites for example. Subject browsing on data is also a service which is provided by half of the cases as a search service homing in on the disciplinary needs of the researcher. Three cases also reported having customised interfaces to encourage more deposit by specific disciplinary groups as is the case for HAL for example who has designed various interfaces with a look and feel of certain disciplinary groups. Half of the cases also reported pushing out

[8] Koninklijke Bibliotheek (National Library of the Netherlands): http://www.kb.nl/index-en.html
[9] For more information on the E-depot and digital preservation at the National Library of the Netherlands see http://www.kb.nl/dnp/e-depot/e-depot-en.html
[10] INTUTE: http://www.intute.ac.uk/
[11] Google Scholar: http://scholar.google.com/
[12] arXiv.org, Cornell University Library http://arxiv.org/
[13] RePEc (Research Papers in Economics): http://repec.org/

information to the depositor such as providing download statistics on material supplied and RSS feeds on new entries.

Various types of services were identified in the following areas: 1) resource discovery, 2) bibliographic services, 3) marketing, 4) Web 2.0, 5) new data online, 6) preservation, 7) file formatting, 8) statistics, and 9) networking.

1) *Resource discovery.* Services in this area vary from standard ones such as search and browse facilities to others which encourage users to return to repository services such as the Connecting Africa subject portal with a Pick of the Month resource highlighted on the front page of the site where a librarian posts either an article or image. More dynamic services such as "Others who consulted this document consulted ...." or hyper-linking bibliographic references were services mentioned by CERN which provide the end user with increased access to new content. The push technologies mentioned were RSS feeds and disseminating content out to other information services be they disciplinary like arXiv.org or more generic online search services such as Google or Google Scholar. Search services provided in this area include views onto repository content which can be chosen by subject discipline, keyword, and author for example. Lists are also generated by some of the repository services like HAL and Cream of Science following personal preferences such as publication type, domain or organisation. In the case of HAL, subject domain searches for example, can clearly identify collaboration between organisations and thereby further highlight French research networks. Customised deposit interfaces also exist to encourage either research group deployment on one level or organisational participation on another as is the case for HAL.

2) *Bibliographic services*. These services are provided to help save the researcher time on administering personal CVs or publication lists and to provide easier access to full text through such lists. Repositories studied provide automated publication lists with links to academic output. CVs are also updated with archive-generated content. Bibliographic lists can also be exported in some cases as with Southampton or HAL. Publications deposited in the repository were also in one case (ASC, Connecting Africa[14]) additionally archived on a CD-Rom and given to the researcher as a digital collection of their work.

3) *Marketing.* ECS Southampton pointed out the importance of marketing recent repository deposits. For example, at ECS most recent publications are showcased on plasma screens at the entrance of the ECS School publicising recent research results. Otherwise, websites can be fed with such IR content. Marketing the work deposited in your repository will help show the currency of information in the repository, will showcase recent work and can ultimately stimulate others to deposit.

[14] ASC - The African Studies Centre: http://www.ascleiden.nl/

4) *Web 2.0.* At the time of interview, few cases reported services in this area. One service was mentioned by CERN which addressed more interactive social networking and interactive information exchange: the addition of commentaries to documents bringing new dynamic content to the researcher.

5) *New content online.* Retro-digitisation is the means to provide new digital access to existing content by scanning material, OCRing, etc. This effort can contribute to providing more online access to historical academic output or to certain types of publications which are not yet available online such as books or their chapters.

6) *Preservation.* Archiving a copy of research material in a repository is a back-up to the original work and this was mentioned as a service. For future longer or mid-term preservation Southampton is cataloguing data formats. Some see it as a prime goal to take on long term preservation such as HAL whereas others see this as a national responsibility.

7) *File formatting.* Repositories provide access to PDF files and therefore generate PDF files on behalf of some authors. Other document formatting can be carried out by repository staff for the benefit of researchers in certain disciplines such as LaTeX which was reported by HAL.

8) *Statistics.* Usage statistics, i.e. download logs of material accessed via the repository, is a service which can feed back new information on the world-wide use or rather access of deposited material. Statistics are provided to its researchers by Minho, Southampton and HAL. This is a further means to visualise the impact of an author's research.

9) *Networking.* Repositories can serve to extend personal networks by bringing content and persons together in subject-community portals as is the case with Connecting Africa.

6. The role of the discipline

It is essential to address the specifics of the disciplines and communities if we are to better guarantee consistent content deployment in OAI-PMH archives. By targeting their needs and adapting to them as much as possible, local content deployment aims will be more manageable.

Understanding the discipline, i.e. how researchers are evaluated, how they organise their research output and where they publish, is essential before repository managers approach researchers to collaborate in repository deposit. Understanding work processes and how material is stored and organised in a particular area can result in repository managers successfully identifying ways to engage with researchers by synergising with current efforts such as the Research Assessment Exercise (UK) for example. Repositories need to work closely together with Current Research Information Systems (CRIS) so that material is deposited once and once only, and

utilised for various purposes to maximise efficiency and to lessen the burden on administration for the researcher. Of the six cases studied, five either have a CRIS-like function or exchange data.

Collection development choices for the repository must depend on the nature of current and future academic domain outputs. Disciplines with a tradition of self-archiving can play an important role in identifying areas where online material is already available, which can then be harvested by the repository. This is the case for CERN and its particle physics community, where CERN "re-claims" about 30% of its academic output this way by harvesting both metadata and full text (in selected cases) from about 100 sources. Alternatively, disciplines can highlight material which is currently difficult to access online as in the areas of the Social Sciences and the Humanities. Repository managers can focus on unlocking new content in these cases and providing new opportunities for increased visibility. For this reason, it is important to give departments some autonomy and decision-making powers to determine what types of material should be collected and disseminated. This is a practice adopted by CERN, Minho and Southampton for example. This can include both material which is regularly used for research as well as material which is more difficult to access such as datasets. Publication channels vary from one discipline to another. For example, books and monographs are a popular medium for Law and the Humanities whereas in the area of Economics, working papers and journal articles are of greater importance. In the area of Computer Sciences, conference papers and journal articles are key.

More visibility means increasing the chances of access and use of research results, with the potential to improve on citation numbers and researcher reputation. The issue of the lack of visibility for certain disciplines is a larger issue for some subject areas. Disciplines with less national or international visibility will profit from increased online impact as and when more material is deposited via repositories. Certain disciplines such as particle physics already have a tradition of over 10 years in self-archiving material. This makes deposit in an institutional repository more of a challenge because researchers are often unwilling to deposit material twice. However, all cases reported visibility as a reason for researchers to deposit, including CERN and its particle physicists. Another aspect of importance is the importance of the IR to the institution, its growth and impact.In 2006 Minho's Vice-Dean of Research publicly announced that Minho had been chosen as one of MIT's partners in several research and teaching projects partly due to the increased visibility of Minho's work through its IR.

7. Selected guidelines

Guidelines to support repository managers in their quest to populate repositories are the result of the research undertaken with all six cases. They draw on the critical success factors and inhibitors highlighted by the cases studied. Guidelines can be grouped into three areas: pointers for better take-up by the research community, guidelines for the better deployment of content, and pointers for ensuring professionalism in the process.

7.1.   For better take-up from the research community

When organising your repository, it is vital to address the diversity between disciplines and their needs, realising that there is no 'one size fits all' solution for incentives to deposit material. Bio-chemists, economists, physicists, geographers, historians all follow different research work processes, the way they organise their research and their networks differ, and the way they disseminate that output varies. These aspects need to be considered when creating advocacy programmes, designing deposit or end user information retrieval service interfaces, thinking of content ingest methods, and so on. Inform and involve your researchers from the outset and ideally involve them in the decision-making process of determining what type of content should be deposited, as is the case at Minho and Southampton for example.

Ownership at the place of production, i.e. research department, is advantageous in making a sustainable service and archive in the future. Making faculty or departments responsible for deposit, and creating sub-archives or deposit interfaces designed for them, can make them feel accountable for the work done and can further stimulate deposit. Get departments involved in deciding collection development policy for the repository for their discipline and in contributing to other policy decisions. This can encourage the embedding of repository activity into the work processes of the department. Departmental representatives can then liaise with repository staff on technical or IPR issues when needed. Minho University has taken this approach with, as of late 2006, 26 out of 30 research centres contributing to the repository in this way. Bio-engineering at Minho is contributing about 80-90% of its academic output for example.

A thorough communication plan which outlines both internal and external communication channels is important to effectively identify and target all stakeholders and their needs for advocacy purposes. Identify what each group needs to be made aware of, know what their needs and fears are, consider the products needed to advocate your message per group separately, and make concrete time plans. The communication plan also needs constant review as and when stakeholders change or as and when the repository develops. For an example of a well-developed plan, turn to Minho University once again.

As far as the repository message is concerned, be sure to be clear about what open access stands for and what the benefits are to the information provider contributing to repository efforts, before addressing researchers. Let open access be the selling point but also above all the added value services which answer researcher needs, emphasises HAL. Create services which save the researcher time on his/her administrative activities. Showcase your efforts and achievements by publicising recent results for example on departmental websites and other media such as plasma screens showing recent publications as at ECS, Southampton for example, and thereby motivate others to deposit.

7.2.   For more deployment

Various aspects are important to help in the deployment of content for Institutional Repositories. It is important for two reasons to take on an active role in relating repository work to the area of information retrieval and discovery services worldwide both generic and specific to the disciplines you serve. Firstly, in order cost-effectively to aggregate existing content, know where your researcher's work is stored: in local archives, in disciplinary archives, or other larger regional ones such as data archives. Use these addresses to *pull* material into your own archive, making agreements with information providers on metadata *and* full text aggregation. CERN for example, made agreements with arxiv.org to harvest both metadata and full text from their researchers who can prefer to deposit in a system they know and use well, as opposed to an institutional archive.

In addition, know the services and search engines which provide further accessibility to your researcher's work such as Google, Google Scholar, or disciplinary services. Know where the researcher searches for others' work and where he/she wants to be found. It is then for you to have additional repository services in place to *push* out repository content to these locations. This will increase the visibility of your authors saving them time on disseminating their work. *Push* out content to the world research community to show your commitment to increasing the impact of your researcher's work. HAL and others push out material to RePEc for example, and others such as Minho and Southampton have looked at ways of maximising the visibility of their IR content in Google Scholar for example.

7.3.   For professionalism

To summarise the above, any professional repository or repository-based service should strive for cost-effectiveness by 1) analysing work processes to be able to target users well, 2) considering harvesting existing content and using web services if this means a more economic way of aggregating content be this metadata or full text, 3) once an initial proof of concept exists of the archive, encourage departmental deposit of content with mediation for reasons of ownership and sustainability, and 4) use knowledge exchange to evaluate current efforts and build upon them based on lessons learnt.

Repository managers should do their utmost to ensure one-time deposit of academic output as far as possible for their authors. For this reason, linking with Current Research Information System (CRIS) must be a given in future. Five out of the six cases had connections to their CRIS, or had a dual function as repository and CRIS for the institution. This is an efficient way of using resources, where material can be deposited once and used for annual reports, evaluation exercises as the Research Assessment Exercise (RAE) as at Southampton, or to feed repository services.

All cases studied specified the importance of knowledge exchange within local, regional, national and international networks and how they can play an important part in contributing to more efficient, effective and state-of-the-art repository results. For example, the implementation of that knowledge can contribute to the development of

policy by following policy developments world-wide, being aware of other services under development and adapting those to suit local needs. All partners reported that networking was important for their work. Last but not least, by being aware of repository changes and networking, repository personnel have the opportunity to advance which in turn affects repository standing.

One of the largest barriers to repository deposit is the fear of jeopardising relations with publishers by storing material open access. The repository team therefore needs to provide sound professional support and advice on IPR issues where researchers are informed of the opportunities open to them. Ideally, library staff are trained in the area of digital rights management and can communicate opportunities open to researchers succinctly. They will also utilise tools such as SHERPA/RoMEO.[15] For specific legal questions, a legal expert should be on call.

8. Conclusions

Measuring the success of your repository is not based on numbers. How far have you reached your objectives with your repository or service? The profile of your institution or service, and the disciplines it serves, will determine your scope, the full text numbers you hope to achieve and the services built around your archive. A number of factors influence the stimulation of the population of repositories. Policy and organisational issues play a role as do advocacy mechanisms. Good, active user relations are crucial to link into the existing work processes of the author, making the threshold to collaborate as low as possible. Synergising with existing activities such as evaluation or reporting mechanisms, i.e. a CRIS, or identifying opportunities through service development to give back time to the researcher for actual research can better guarantee repository deposit and help forge longer-term relations between library and client.

Further knowledge can be gained on the six repository cases by reading the individual case write-ups listed at:
 http://www.tilburguniversity.nl/services/lis/driver-population.html

---

[15] SHERPA/RoMEO: http://www.sherpa.ac.uk/romeo.php