Beyond institutional repositories

*Laurent Romary, INRIA (Gemo research group) & Humboldt Universität zu Berlin (Institut für Deutsche Sprache und Linguistik)
laurent.romary@loria.fr
http://www.inria.fr/recherche/equipes/gemo.fr.html
Unter den Linden 6, 10099 Berlin - http://www.linguistik.hu-berlin.de/

Chris Armbruster, Max Planck Society (Max Planck Digital Library)
armbruster@mpdl.mpg.de
Invalidenstrasse 35, D-10115 Berlin - www.mpdl.mpg.de
Executive Director, Research Network 1989 - www.cee-socialscience.net/1989

Abstract
The current system of so-called institutional repositories, even if it has been a sensible response at an earlier stage, may not answer the needs of the scholarly community, scientific communication and accompanied stakeholders in a sustainable way. However, having a robust repository infrastructure is essential to academic work. Yet, current institutional solutions, even when networked in a country or across Europe, have largely failed to deliver. Consequently, a new path for a more robust infrastructure and larger repositories is explored to create superior services that support the academy. A future organisation of publication repositories is advocated that is based upon macroscopic academic settings providing a critical mass of interest as well as organisational coherence. Such a macro-unit may be geographical (a coherent national scheme), institutional (a large research organisation or a consortium thereof) or thematic (a specific research field organising itself in the domain of publication repositories).
The argument proceeds as follows: firstly, while institutional open access mandates have brought some content into open access, the important mandates are those of the funders and these are best supported by a single infrastructure and large repositories, which incidentally enhances the value of the collection (while a transfer to institutional repositories would diminish the value). Secondly, we compare and contrast a system based on central research publication repositories with the notion of a network of institutional repositories to illustrate that across central dimensions of any repository solution the institutional model is more cumbersome and less likely to achieve a high level of service. Next, three key functions of publication repositories are reconsidered, namely a) the fast and wide dissemination of results; b) the preservation of the record; and c) digital curation for dissemination and preservation. Fourth, repositories and their ecologies are explored with the overriding aim of enhancing content and enhancing usage. Fifth, a target scheme is sketched, including some examples. In closing, a look at the evolutionary road ahead is offered.

Keywords
Scientific information, publication repositories, institutional repositories, digital libraries, research infrastructure, deposit mandate, interoperability, open access

Publication repositories at the crossroads

A series of developments over the past two decades has shaped scientific communication such that parallel to the long-standing tradition of scientific publishing, online environments have been set up to provide fast, wide and free access to content by means of *publication repositories*. This paper posits a crossroads for publication repositories, tracing contextual factors that explain why and how it is necessary for us to reconsider the basic parameters of how publication repositories should evolve further.[1]

The current system of so-called institutional repositories, even if it has been a sensible response at an earlier stage, may not answer the needs of the scholarly community, scientific communication and accompanied stakeholders in a sustainable way. However, having a robust repository infrastructure is essential to academic work. Yet, current institutional solutions, even when networked in a country or across Europe, have largely failed to deliver (cf. Basefsky 2009). Consequently, a new path for a more robust infrastructure and larger repositories is explored to create superior services that support the academy. A future organisation of publication repositories is advocated that is based upon macroscopic academic settings providing a critical mass of interest as well as organisational coherence. Such a macro-unit may be geographical (a coherent national scheme), institutional (a large research organisation or a consortium thereof) or thematic (a specific research field organising itself in the domain of publication repositories).

We are concerned with crossroads. Therefore this paper will neither trace the history of online scientific communication nor consider the debate on open access, except when it is directly relevant to the argument (for background information consult Armbruster 2007, 2008a, 2008b and 2008d). To substantiate the claim that it would be wise to reconsider the parameters of the publication repository infrastructure, we proceed as follows. Firstly, while institutional open access mandates have brought some content into open access, the important mandates are those of the funders and these are best supported by a single infrastructure and large repositories, which incidentally enhances the value of the collection (while a transfer to institutional repositories would diminish the value). Secondly, we compare and contrast a system based on central research publication repositories with the notion of a network of institutional repositories to illustrate that across central dimensions of any repository solution the institutional model is more cumbersome and less likely to achieve a high level of service. Next, three key functions of publication repositories are reconsidered, namely a) the fast and wide dissemination of results; b) the preservation of the record; and c) digital curation for dissemination and preservation. Fourth, repositories and their ecologies are explored with the overriding aim of enhancing content and enhancing usage. Fifth, a target scheme is sketched, including some examples. In closing, a look at the evolutionary road ahead is offered.

[1] For input, comment and criticism we thank Richard Boulderstone, Foudil Bretel, Christoph Bruch, Natasa Bulatovic, Lee-Ann Coleman, Malte Dreyer, Adam Farquhar, Laurent Guillope, Hannah Jenkins, Jacques Millet, Alain Monteil, Uwe Müller, Neil Sandford and Ulla Tschida.

Deposit mandates: towards a single repository, common format and integrated services?

Deposit mandates are seen as the most important route to enabling more open access to scientific knowledge. Funders' deposit mandates seem particularly important because they target high quality research output, thus setting an example for scientific communities as well as academic institutions. When the National Institute of Health (NIH) implemented its public access policy in 2008, deposit in Pub Med Central (PMC) was mandated, a digital archive developed and supported by the NIH as a single repository with a common format.[2] Earlier, in 2006, several UK research funders in the life sciences had also opted to create a single repository with a common format, UK PMC (linked to PMC), to implement their deposit mandates.[3]

While it could be argued that subject-oriented funders would favour subject-based repositories, it is nevertheless remarkable that they would opt for a single, shared repository. Moreover, the European Research Council (ERC) and the European Commission (EC, as research funder), which fund across disciplines and countries, are also opting to create a single repository. Initially, the European Commission had favoured institutional repositories, later advocating deposit in institutional and subject-based repositories alike. However, once deposit mandates were being implemented, it became apparent that a high-quality repository service is required, which is achieved most likely by a single repository with a common format.[4]

Single repositories are providing distinct value to funders, such as helping them to manage their relations with grantees (e.g. reporting), improving internal knowledge management (e.g. portfolio management) and providing the public with a comprehensive overview of research results achieved (e.g. accountability). Moreover, the common format helps the repository manager to develop services that are of value to grantees (authors) and users alike, such as citation services tracking the impact of a publication dynamically. Moreover, single repositories, which typically are large and shared, are accomplishing a growing volume of direct publisher deposit.

A difference must be observed between posting a working paper or preprint and depositing a final published result. Authors will post a preprint to claim priority and inform colleagues, but for the deposit of a final published result the author is neither the most interested nor the right agent, as professional intervention by publishers and librarians is required. Experience with large-scale deposit by publishers and deposit assisted by librarians imparts two lessons. Firstly, the notion that a deposit mandate would nudge authors to do a few more keystrokes to self-archive a final version is principally mistaken. Providing open access to final published results, whether the author's final manuscript or the publisher's version, requires quality checks (version control, metadata) and long-term solutions (archiving, access) that mean that librarians

---

[2] http://publicaccess.nih.gov/

[3] http://ukpmc.ac.uk/

[4] European Commission (2008) Open Access Pilot in the European Commission's Seventh Research Framework Programme (FP7). Special clause 39 on Open Access. http://ec.europa.eu/research/science-society/document_library/pdf_06/ec-open-access-pilot-ppt_en.pdf - leading to the call FP7-Infrastructures-2009-1

and publishers are much better placed to handle the deposit. If this is so, then a second lesson follows, namely that a large-scale solution is the economically more efficient way of organising the deposit of final research results.

By contrast, there exist well over one thousand institutional repositories, the majority of which hold very little content. Further still, these repositories have no common format. The one integrated service operational for many years, a protocol for metadata harvesting, does not guarantee the most basic service that repositories must offer: search and find. A critical experiment for high-quality self-archived articles has clear results (Bergstrom, Lavaty 2007). First, it was established that for thirty-three key economic journals (of a given year), ninety percent of articles in the most-cited journals had been self-archived and about fifty percent of articles in less-cited journals were available. Second, all freely available articles could be found using the Google search engine. However, when using OAIster to search (institutional) repositories (based on the Open Archives Initiative – Protocol for Metadata Harvesting), only one quarter of the articles were found.

Institutional repositories initially may have seemed like a good way to enhance the visibility of research results produced at any institution. Also, they seemingly reflect the grassroots nature of the open access movement and the initiative of individuals, departments and libraries. Yet, a glance at a ranking of repositories for visibility, quality and available items reveals a rather short head of fairly large repositories based on research publications (high-quality preprints included).[5] The standard argument of proponents of the institutional solution is that deposit mandates will help to fill (and institutional repositories backed by a mandate do reasonably well in the ranking), but this still leaves open the issues of a common format and integrated services. Moreover, problems of coordination and cost would seem insurmountable already in the medium term, compounded by the problem of how to guarantee permanent access.

Comparison and contrast: central research publication repositories versus a network of institutional repositories

Central research publication repositories are very attractive for their quality. Research funding is usually highly selective and the proposal must demonstrate originality and novelty. Selected research proposals are adequately funded. The publishable results are submitted to the best journals. In the process, the research design and results often are vetted multiple times by peer review. A central research publication repository is therefore likely to hold the best, most original and newest research. Moreover, they allow for portfolio management. Funders and institutions easily may track results through publications (and data sets) while users have an information tool to look up grant holders and research results.

An institutional repository contains the various outputs of the institution. While research results are important among these outputs, so are works of qualification or teaching and learning materials. If the repository captures the whole output, it is both a library and a showcase. It is a library in that it holds the collection. It is a showcase because the

[5] Webometrics 2009 Ranking of World Repositories http://repositories.webometrics.info/index.html

online open access display and availability of the collection may serve to impress and connect, for example, with alumni of the institution or the colleagues of researchers. Moreover, such an institutional repository could have an important function in regional development. It allows firms, public bodies and civil society organisations to immediately understand what kind of expertise is locally available. Institutional repositories may also support internal and external assessment as well as strategic planning.

A more systematic comparison may be undertaken across key dimensions. Taking into account, roughly, what is known about institutional repositories on the one hand and central research publication repositories (e.g. Arxiv, SSRN, HAL, RePEc etc.) on the other hand, the following picture emerges.

	Central repository	**Institutional repository**
Deposit	Submission system that seeks to maximize publisher deposit and assisted deposit, backed by publishers and librarians.	Self-archiving, requiring the author(s) to submit (including metadata), possibly assisted by repository staff.
Quality	Final publication primarily, peer reviewed, with quality imprimatur, possibly with supplementary material such as data – often based on results of funded research, with earlier peer review of proposals.	Wealth of material, much qualification work, final publications only part of the collection.
Visibility	Coherent collection, alert services to the research community. High direct value to active researchers.	Reliance either on generic search engine or, else, on interoperability (federated search or portal).
Access	One-stop shop of research results with additional value from overlay services (e.g. metrics) and re-use potential (e.g. mining).	Some insight into activity at institution, else reliance on federated services.
Standards	Unified and high standards for services, access and preservation may be set; any correction of standards is easy.	Standards must be negotiated, agreed and implemented; any change is subject to the same procedure.
Preservation	A single solution for preservation and migration may be adopted.	Preservation must be achieved at each site; else content must be migrated to a central archive.
Cost	Calculable as a (small) percentage of research funding and expenditure.	Additional expense to the institution, which may be distributed by relying on labour of institution's members.

Table 1: Contrast of repository models

Two observations would seem to follow. Firstly, a central solution has some distinct advantages. The most important is that any central repository is part of the natural environment of the scientist in a way that the institutional repository will never be. A connected additional advantage is that unified and high standards support the continuous improvement of services. All this may be achieved while reducing the burden for the author. It would also seem that maintenance and preservation may be managed more efficiently with a single repository, making this the more easily sustainable solution. Secondly, it emerges that institutional repositories and their networking is a counter-intuitive solution. Going against the grain sometimes is successful and the proponents of institutional repositories did believe that this was the best way to proceed from spontaneous self-archiving (estimated to be limited to about 15% of research results) to universal open access. However, if one compares the content available, the level of service and the potential of the two alternative models, then it emerges that institutional repositories have a future that is strictly limited to cases in which institutions have the resources and consider it a priority to host an institutional showcase.

Publication repositories – the big picture

After arguing that central publication repositories perform better across key dimensions, we move to consider important functions that repositories have for both researchers and research institutions.[6] As depicted in Figure 1, repositories need to be able to fulfil the following functions:

- Provide an infrastructure whereby scientific information is widely disseminated and accessible to researchers and the public alike;
- Establish a reliable environment that certifies information with regard to the depositor, the reliability of the information being deposited and the time at which the information has been deposited;
- Guaranty long-term availability, meaning that repositories are also libraries.

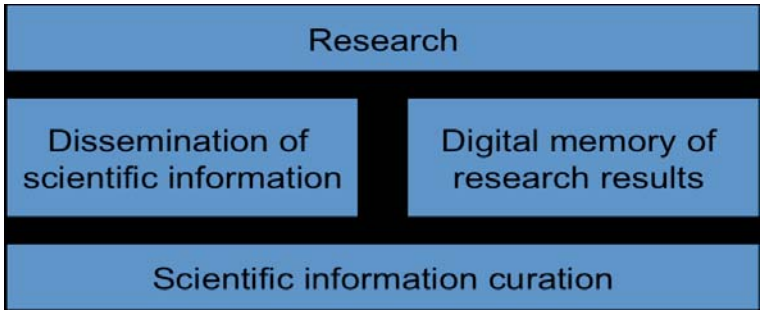


Figure 1: The big picture

[6] Research institution refers to institutions with research funding and/or research performing aspects.

Fast and wide dissemination of scientific information

The first and foremost reason why a researcher may be interested in having a publication repository at hand is because it allows her or him to do more efficiently what the job is all about, namely communicating ideas and results. As soon as electronic communication became available, researchers have widely used email, ftp sites and, ultimately, web pages to disseminate their publications or drafts thereof. What they basically expect from such environments is the capacity to have their work reach actual or potential colleagues immediately.

In fact, researchers are reluctant to invest more than what is strictly necessary in these dissemination activities. We cannot expect for instance to have them accurately key in precise metadata such as the exact journal title, publisher, pagination, and the like. Besides, the document that is posted or transmitted is the one at hand and only very seldom will efforts be made in identifying which format would ensure the greatest legibility of the work.

Researchers' interest has led to the development of central repositories, mostly articulated around the dissemination of pre-prints, whose function is to act as a direct gateway between researchers and their colleagues. This view of repositories as direct scientific communication tool has been the source of development of specific researcher oriented services such as prior scientific validation, alerts, thematic organisation of content, together with minimalistic metadata profiles, allowing further harvesting by portals. Paradigmatic examples of such a repository framework are Arxiv, RePEc or SSRN, which in some domains have become major content holders and service providers. HAL, as a national infrastructure, was initially conceived as a researcher's tool mirroring the functionalities of Arxiv.[7]

Another important aspect of the dissemination philosophy is that researchers are naturally tempted to associate to a publication corresponding accompanying material that has either been material for the research or which has been published in combination with the paper. Such material may cover research notes, laboratory notebooks, slides presented at a conference, various illustrations or graphics, software and reports of all kinds. The potential complexity of accompanying material, from the point of view of their storing and documentation, may bring in a supplementary order of complexity, and potentially of fuzziness in the context of publication repositories.

It can be observed that even if researchers are at the centre of the dissemination activities, research institutions may also be interested in having scientific results widely disseminated. When they do so, they express further requirements, as we shall see in the following section.

A digital memory of research results

A complementary view on publication repositories is to consider them as a preservation area of the scientific production associated with an individual or a group. Even if mainly

[7] For further information on repositories we suggest to use as a starting point http://repositories.webometrics.info/index.html

associated with the vision that a research institution would have on publication repositories, it should be noted that individuals also consider a publication repository a reliable place where their publications will be preserved and where they themselves will retrieve their various outputs over the years. Furthermore, this view relates also to the political expectations of a society towards its research institutions, in that they are able to preserve, as would be done for any other kind of cultural heritage asset, a trace of their activities and discoveries. What is expected is the capacity to manage publication material in such a way that, on the one hand, at any time in the future this material will be retrievable and legible, and, on the other hand, this material represents an accurate and comprehensive picture of the actual research output.

The issue of retrievability can be linked to the capacity to provide long-term archiving access — in a broad sense — to the corresponding content. By long-term archiving we mean here not only the capacity to store data in a reliable way at bit level, so that basically no information loss would occur, but also the capacity to document the data in such a way that any document is associated to enough descriptors to make it uniquely recognisable by search or navigation means. Any simplification in the metadata description process that would lead to a blurred identification of digital items may lead to a repository becoming a digital cemetery where information is lost forever.

Legibility has to do with the capacity to get access to the informational content of a document independently of the technology that has been used for its creation. This requires either on the part of the depositor or on the part of the data curation component of the repository itself to strive for innocuous data formats that are based, as much as one can, on official or open standards. This may require, for instance, the allocation of specific manpower dedicated to the transformation process from proprietary format to actual well-defined standards.[8]

Permanent access imposes far more constraints on the way information is to be managed within repositories and can be seen as quite an overhead for anyone wanting a simple dissemination service. Still, providing such archival facilities on top of a simple dissemination process with basically no extra cost for the individual user (i.e. the scientist), leads to a coherent picture where the two views form a solid background for justifying that such a service as a publication repository should systematically be provided to scientific communities.

Clarification of this second function underscores that the author is not the best or right agent for the deposit of final published material (and its supplements). Moreover, a multitude of institutional repositories will always be able to achieve this function only a great overall cost, because it would have to be undertaken at every site. Of course, individual repositories could outsource preservation and permanent access to a central archive, but that would only confirm that central repositories are the more sustainable solution.

[8] For a discussion of issues related to the definition of an XML based format for journal papers see, for instance, Holmes and Romary, 2009.

Digital curation by research libraries

Librarians can be positioned in the big picture as crucial in supporting fast and wide dissemination as well as permanent access. Indeed, we consider that publication repositories are embryonic to the wider notion of a digital library, which should not be decoupled from the current research library infrastructure, but, to the contrary, be seen as natural target role for them.

The validation and enrichment of metadata would be one basic activity to enable enhanced services that are both fast and lasting. We can identify some core domains of intervention by order of importance:

- Bibliographical information: this is probably the most important domain since it fulfils both the expectations of scientists to be quoted adequately and that of institutions to have precise information about the actual scientific production. Among such information is the systematic proofing of article title, author list, journal title, imprint information (volume, issue, pagination) and core identifiers such as ISSN and DOI;

- Identity and affiliation: this comprises the disambiguation of names and the various levels of institutional linking that bear upon an author, such as research team, laboratory or encompassing institution. This should reflect in particular the complexity of multiple affiliations so that any attribution of the work can be made with accuracy. The proper management of such information is indeed a key issue to ensure trust on the part of research institutions;

- Keywords: these may be provided by the author, but they should be produced preferably by librarians, or at least checked, e.g. in relation to reference vocabularies such as the MeSH[9] or centralised databases such as Termsciences[10] (see Khayari et al., 2006). A well-defined editorial policy in this respect facilitates the definition of coherent views on repository content.

Publication repositories and their ecology

Before we can think any further on a better organisation of publication repositories we suggest to take another look at repositories and their ecology, as depicted in Figure 2, which combines roles allocated to a repository with possible factors that these roles may depend on. We aim to show that none of those factors should be considered in isolation, but taken together. Hence the following discussion remains at a high level and does not delve into the large amounts of literature available.

[9] Medical Subject Headings - http://www.nlm.nih.gov/mesh/

[10] http://www.termsciences.fr/

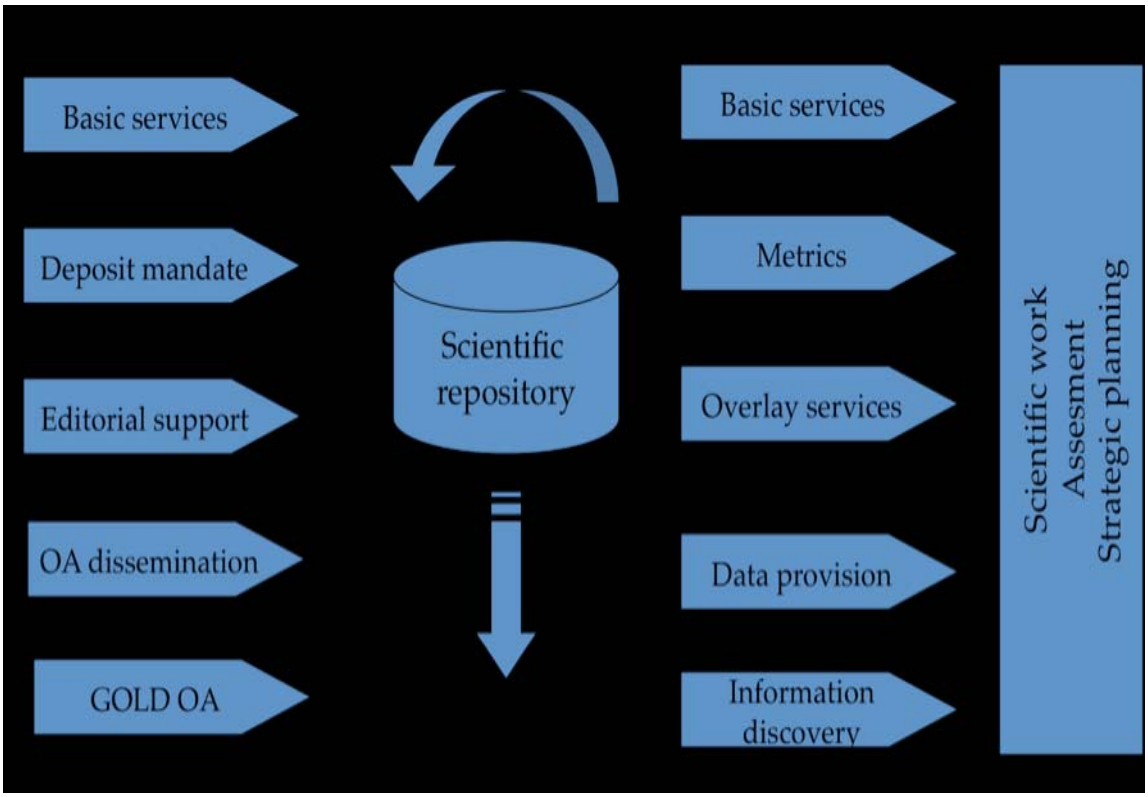


Figure 2: Repositories and their ecology

The core benefits of a publication repository

The three main attractors that, when put together, explain the variety of perceptions on publication repositories and thus the necessary degree of complexity of any appropriate organisational structure, are detailed as follows (see right hand-side of Figure 2).

Scientific work

A publication repository should be first and foremost a tool at the service of researchers. Their own output should be easily available to them as a personal archive and, in a second instance, as a communication tool towards their scientific community. In a way, they should not have to be acquainted with the open access concept to just adhere to the idea that a publication repository can become an essential aspect of the management of their research process. In this respect, the scientist should feel at ease to manage his own editing workflow, starting with the first research notes until the final publication, and encompassing additional material (illustration, data excerpts), within one single environment corresponding to his workspace within the repository.

Assessment

One cannot escape mentioning the importance of the role of repositories as an underlying source of information for the assessment of researchers and academic institution. Publications can indeed be considered as the core of any assessment campaign, both from the point of view of the actual content and their visibility (e.g.

citation metrics). Still, it is obviously a touchy issue and once again, the focus has to be on the scientist, so that the repository is a facilitator for her or him to produce the information required for assessment. Researchers should indeed select the collection of papers they want to make visible, tune the way the corresponding information is made accessible and exported, and have the capacity to integrate seamlessly such facilities with other reporting tools.

Strategic planning

A complementary issue to that of assessment is for research organisations to get an overview of their scientific impact within their various fields of activity. This aspect relates to the necessary strategic planning that all research organisations have to perform and depends on having a reliable and comprehensive view on where, how much and with which visibility researchers affiliated with the institution have published. The baseline for such an activity is usually to rely on external publication databases such as the Web of Science, from which information may be extracted according to factors such as research domains or institutional belonging. However, this appears to be a hazardous path to follow for two reasons:

- The thematic coverage of such databases is uneven and, as pointed out by bibliometric studies (cf. Butler; 2006), does not take into account the specificities of domain where, for instance, publication occurs frequently in conferences and workshops;

- The treatment of affiliations is by far too shallow to allow a precise selection of the material relevant to a specific institution. Problems may range from ambiguous names to incomplete coverage of multiple affiliations, which require enormous additional manual work as well as caution in the interpretation of results.

In both cases, publication repositories, through their potential capacity of providing both a wide coverage of an organisation's production as well as enriched meta-data, should play a central role in consolidating bibliometric data. In turn, when properly handled, this factor may be an important additional reason for individuals and institution to adopt a repository.

Basic services to enhance content and usage

The main priority for a repository should be to stimulate scientists' interest through the provision of core and simple services corresponding to the upper layer of Figure 2. Such services may be based on the following list:

- Easy submission: the deposit process should be made as simple as possible through simplified deposit forms with reduced mandatory fields, direct connection to major subject based repositories (Arxiv, PMC) where the paper may already have been referenced, availability of a "favourite co-authors" list, and eventually automatic metadata extraction from the document;

- Stable reference: the scientist should trust the repository in its capacity to both allow her or him to point to (persistent identifier scheme) and retrieve (persistent archiving) publications in a durable manner. This may allow the researcher to

use repositories for any reference to her or his work or to inform a colleague about it;

- Publication list: one of the main ways, beyond simple Google-based searches, users may want to retrieve information in a publication repository is to be able to extract a list of published material, with various types of orderings (author, dates), organisations (document type, year of publication) or output formats (formatted vs. structured output);

- Web pages: this is usually a by-product of the publication list service, but requires specific attention since this will be a component of the first line presentation and communication tool of potential users at various levels (authors, departments, institutions). It should allow a high level of customisation while offering ready-to-use layouts. When linked to a Content Management System, an automated generation of web pages can be seamlessly integrated with a laboratory web site.[11]

Editorial support

The three main areas where editorial support (through research librarians) may provide benefits to a publication repository and its users are the identification of content, support of deposit, and enrichment of content.

The identification of content is a pro-active endeavour to work closely with the scientist so that they are aware of the existence of the repository and are encouraged to signal publications ready for deposit. More generally, it means to trace publishing activity relevant to the repository. Identified content needs to be deposited and this may be supported in a number of ways, not only by organising (automatic) publisher deposit and setting up an easy manuscript submission system (assisted deposit) but also by providing service that ensure visibility and dissemination to the relevant communities (e.g. exchange with other repositories of a stable version). More generally, thematic collections are likely to be the focus of dissemination activities. Central repositories (e.g. Arxiv, SSRN or RePEc) demonstrate that pro-active dissemination by notification mechanisms is highly appreciated by researchers and turns the repository into a valuable resource.

Beyond open access – collaboration schemes with publishers

For repositories to work across the disciplines, a strong open access policy is an important factor for the success, visibility and use of a publication repository. Ideally, the scientific publishing system would be switched to a non-exclusive licensing system to enhance market competition (Armbruster 2008a). Still, one may consider such a policy from a systemic point of view so that it relates to the functions identified in the big picture. A variety of policies may contribute to the development of a publication repository (Romary, 2007), including agreements with publishers to provide open

[11] See for instance the publication pages of the Max Planck Institute of Psycholinguistic under: http://www.mpi.nl, generated from eSciDoc content.

access to the final published version,[12] and also more standard subscription agreements. These can be summarized as follows:

- Gold open access schemes allow publishers' versions to be transferred to repositories, and made freely accessible, also reporting usage and tracking citations;

- When items are not open access, every effort should be made to reach an agreement whereby publishers' version are at least deposited in a "dark archive", allowing institutions to keep a complete record of their scientific production;

- The lowest level of agreement, nevertheless essential, is to bring publishers to provide precise metadata profiles for all publications relevant to the research institution. This allows publication repository to consolidate precise documentation related to the paper, and conversely gives publishers the assurance that, for instance, backward links to their online services (through DOIs) are systematically indicated.

Providing useful and reliable secondary data

The experience gained from existing central repositories has shown that they are likely to be the most accurate place where reliable metadata, in particular in the domain of authors' affiliation, may be kept. As a matter of fact, when the author sees the repository as a research tool and enough editorial backing is provided, she or he is naturally tempted to detect inconsistencies, e.g. in affiliation. If this issue is taken seriously enough when deploying a publication repository, we can expect that the quality level of the corresponding metadata shall reach higher standards than those observed in third party databases such as the Web of Science, Scopus, or Google Scholar, which usually take the paper version of the articles as their primary information source (Armbruster 2008c).

A more prospective aspect on which publication repositories may provide specific information relation to scholarly publishing is that of access and usage. Indeed, the accessibility of content may be subjected to evaluation (e.g. the Webometrics ranking of repositories) and repositories can gather and consolidate usage information concerning hits and downloads to their content, whether this is metadata or full-text access. Such information is useful to the scientist, as it allows her or him to see how much various publications are searched for. It is also a source of potential indicators about the dissemination role of the repository and, furthermore, a potential contribution to the assessment of scientific content. Even if there exists quite enormous activity in defining possible counting schemes for repository access (see for instance Brody et al., 2009), the reliability of the corresponding information is essentially organisational. The first factor is that of the global visibility of the repository proper whereby it is actually indexed and linked by other information sources (e.g. Google scholar or Scientific Commons). Second, it is essential, for published material to be able to consolidate

[12] See Poynder (2009) for a wider discussion of various open access schemes. One of the authors of the present paper has designed the open access publishing and transfer scheme that was agreed between the Max Planck Society and Springer.

access information with that of publishers. This is beneficial to both sides, since it provides research institution with a real evaluation of the interest raised by each publication, and it is an opportunity for publishers to monitor the usage of repositories, and increase their trust in them, and demonstrate the value of publishing services.[13]

Further services

As can be anticipated, as soon as high quality data is freely available on-line and actually covers a great deal of the existing research production, there is room for developing additional services related to the management of content (bibliographical links, document indexing and classification) or its further exploitation (overlay services).

One important direction to follow, for which a publication repository may be well suited, is to use such an infrastructure as the basis for developing overlay journals, that is peer-reviewed certification processes that use the repository as the locus of deposit of authors' manuscripts. Independently of who is carrying out the certification process - this could be a commercial publisher - such a framework ensures that all versions of the corresponding papers are managed and archived at one single and reliable source.

Target scheme

Following the above reflections, and supported by the experience gained in working with various platforms as well as various institutional environments, an evolutionary and global deployment scheme for publication repositories may be outlined. This would lead to a consolidation of repositories to enhance services for researchers while being more efficient for permanent access. Researchers, funders and policy makers are likely to welcome this consolidation, but it may be expected that institutions and repository managers that have invested into a local solution have developed an interest to maintain their showcase. This is not a problem at all as long as interoperability is given and the content may simultaneously be held in a central repository too. Step by step, the situation will improve.

As outlined, we advocate an organisation of publication repositories based upon critical mass that is geographical, thematic and based on significant deposit mandates from large research funders and institutions. The basic advantages of a coordinated (and sufficiently centralised) approach are the following:

- The environment is well known by the scientists from corresponding scientific communities, whether they are personally supportive of the infrastructure or not. This maximizes chances that the repositories will become part of the natural ecology of researchers.
- Most of the technical services (user IDs, authority lists, workspace facilities, link to existing research databases) can quickly have high operational quality and

[13] The PEER Project has become an arena for cooperation between publishers and repositories, including fresh research on usage, author and user behaviour and the economics of deposit - http://www.peerproject.eu/

follow technological evolution, since development is undertaken by an identified team of developers;
- It brings together research institutions close enough to develop a coherent overall strategy for publication repositories rather then loose time and money in maintaining and networking local environments.

The close coordination of a small number of platforms may deliver a unified and superior level of service by acting at four levels:

- Designing a common portal or, at least, standardized entry points offering the same external visibility;
- Making sure that each platform offers a high level of technical service within a standardised service-oriented architecture, and is quickly able to adopt emerging technologies;
- Ensuring data interoperability and reuse;
- Having a joint long-term archiving strategy ensuring reliable maintenance of the content.

Decentralised editorial support

While we advocate a coordinated repository infrastructure based on large repositories, it would be sensible for the management of data to be positioned as closely as possible to where the research is carried out. By *editorial support* we mean a whole range of personal services that provide support to researchers, teams and laboratories, in order to ensure a wide aggregation of content as well as its enrichment. This basically comprises the various aspects that we identified in the previous sections in relation to digital curation, but we would insist on both the importance of localised services to achieve this and the necessary evolution in the competence profile.

Organisation of editorial support staff is very close to what research library structures have been to date, namely organised around thematic units that are able to keep track of researchers' interests. This facilitates the management of affiliation (knowing the researchers and their collaboration environments), the tracking of content (publication habits) the provision of support to go towards a wider curation of research outputs (knowledge of research methods).

This should be accompanied by some evolution in the information and library science curricula so that the good background in information management is complemented with enough knowledge of digital information methods and technologies. We can anticipate that the term "librarian," in the research environment may soon or later be replaced by that of "digital curator."

On scope

Deposit mandates for research results centre on final publications and data. Indeed, if the scientific record is to become available through repositories, then this is essential. Nevertheless, the value of the repositories should be enhanced by being flexible along several possible axes:

- Drafts vs. published material: to cover both the dissemination and the digital memory views that we identified earlier, repositories should be able to cover documents at an early stage of writing as well as fully published papers. This

- Supplementary materials: it may also be necessary to be flexible as to the kind of item one wants to register. If a researcher's scientific production is to be captured, laboratory notes, case studies, software, workshop papers, tutorials and so on should be recorded if desired;

- Alternative forms of scholarly expression: The exhibition of an urban planner or the simulation of a climate scientist often constitutes not just a supplementary but alternative form of expression, which repositories should account for;

- Overlay services: A variety of services that serve the authors and the users (e.g. access, usage and citation metrics) as well as functions that support web-based scholarly communication (e.g. comments, sharing);

- Bibliographical records: one of the central questions for any repository is whether they should allow depositors to record publications without associating the corresponding full-text. We support the idea that bibliographical records are indeed part of the publication continuum and should be incorporated within the scope of a repository. In many respects, bringing (or mandating) research communities to recording their published production naturally leads them to consider the advantages of also depositing full-texts.

Good management of metadata always allows one to specifically focus on a sub-collection of documents, thus making it quite unnecessary to filter out content at the source of deposit. All in all, publication repositories, seen at the coordinated level we advocate here, are a core component of scientific information.

Towards a more accurate standardisation plan

A more coordinated model enhances the definition of data representation and interoperability standards than any perspective limiting itself to surface harvesting of local publication repositories. The underlying objective is to reach a level of interoperability standards that allows the exchange of the full content of a repository.

From a metadata point of view, the aim is to define representations that form a continuum between the information that document a single paper and that which is used to cite papers. By doing so, we pave the way for a full networking of publication data which in turn can be reused by scientist themselves either to simply get accurate reference lists or to trace citation links between scientific papers.

In the same way, a coherent scheme should be thought of for the representation of full texts, with the perspective to form a continuum at all stages of the publication workflow (cf. Holmes and Romary, 2009), from the writing of a scholarly paper, its dissemination in various formats, and finally its long-term archival in a repository. Such a standardisation activity requires a deep vision about what facilities corresponding formats should offer, as well as a high coordination with repository managers.

Further effort is required for the definition of a whole portfolio of statistical measures that would provide an in-depth picture of access rates to repository context, ranging from basic metadata access to precise tracing of full-text (or contextual) searches.

Sustainability

Over the long term, the most important reasons why we need a more concentrated infrastructure for publication repositories is that of sustainability. The current picture has lead academic institutions to dedicate some small man-power to the initiation of local repositories, without actually integrating this into a master plan to further maintain and improve it. Even more, making a local publication repository a success is, when you actually ask the persons in charge "a considerable effort," that not all universities or academic institution may afford.

By focusing on more centralised technical infrastructures, it is possible to think of business models whereby institutions actually contribute with a small share to an environment which is more likely to be sustainable and, as we already mentioned, bring better services continuously. In turn, this allows academic organisations to put more emphasis on the actual editorial support and local dissemination activities, to ensure that their own production is accurately deposited and preserved in the repository.

Two related examples

To illustrate how the target scheme may be implemented, we provide two major examples of such a coordinated approach. These two repository infrastructures actually encompass many of the technical facilities and user oriented services that we outlined in this paper and can be viewed from two complementary angles:

- From a user's point of view, as the basis for a wider service covering a large community (federation of research institutions, universities);
- From an organisational point of view, as core participants to a reduced cluster of platforms that could offer an efficient and sustainable research infrastructure for publications.

HAL – a national infrastructure

HAL was initially created in 2001 to offer a trustworthy counterpart to Arxiv that could be extended to offer services to the French scientific communities, from both a multi-disciplinary and multi-institutional perspective. It currently contains 120,000 full text documents (among which are 13,000 PhD theses) and has been recognised since 2006 as the reference archive for all major French academic organisations.

The HAL platform is accessed by means of several portals, which allow specific communities (e.g. Human and Social Sciences, information and communication technologies) or institutions (e.g. HAL-INSERM, HAL INRIA, partner universities) to use the system. It also identifies collections for tutorials and PhD thesis. HAL is a full mirror for Arxiv. It also enables conferences to publish their papers online in a dedicated collection.

In its current operational configuration, HAL offers a wide range of functionality:

- Depositor workspace for the management of primary deposit, content enrichment (metadata, complementary documents) and versioning;
- Author and institutional descriptions;
- Basic access statistics (bibliographical entry, full text) for authors;
- Automatic generation of author's collection (e.g. to provide author's web page);
- Multiple export formats (bibtex, endnote, TEI, connection to Crossref);
- Linkage with major subject based repositories (Arxiv, Repec, PMC);
- Simple and complex search (metadata, full text);
- High level interoperability (Web-service API).

Institutions who want to benefit from HAL services may express various kinds of requirements ranging from the simplest usage of HAL to the definition of specific environments. At the simplest level, they may use the generic HAL environment and take part in the quality control associated with each deposit. If need be, they can have a specific layout designed for the presentation of their research output. In specific cases, they may also have their own portal with customized metadata profiles or submission interface. This has been the case for INRIA,[14] which, after five years of operation, covers 32% of INRIA's scientific production across its eight research centres. Editorial support (data validation and enrichment) as well as communication with researchers, is provided by the scientific information networks of the various research centres.

eSciDoc — an open technological platform

eSciDoc is a joint project between the Max Planck Society (MPS) and FIZ Karlsruhe, sponsored by the Federal Ministry of Education and Research (BMBF). Within the MPS, the project falls under the auspices of the Max Planck Digital Library (MPDL), founded in January 2007. The goal of the project is to develop a multidisciplinary, virtual research environment as a part of the government's eScience initiative. The infrastructure and the applications based upon it are made available through an open source license.[15]

eSciDoc is a repository infrastructure platform for management of research publications and data. It is based on Fedora and offers additional services and application solutions to manage various types of scientific information at large, including a publication repository, PubMan.[16] The PubMan solution is being rolled out at the Max Planck Society, starting from May 2009, having already been tested and adopted by six institutes (of the more than seventy). It is intended to replace the former eDoc repository, which was based on a proprietary software platform and appeared to be neither maintainable nor extensible in the mid-term. The current eDoc repository system

[14] http://hal.inria.fr/

[15] http://escidoc.org/

[16] http://colab.mpdl.mpg.de/mediawiki/Portal:PubMan

within the Max Planck Society holds more than 100.000 references and about 30.000 full texts and will be migrated to the new PubMan system.

Being based on a service-oriented architecture, eSciDoc offers the possibility to deploy additional services within the same technical environment, and to compose services external to eSciDoc as part of a standard service offer. Such services can either be specific extensions for special collections,[17] external institutions,[18] or to design scientific data repositories[19] articulated in conjunction with publications.

As a whole eSciDoc combines the strength of a centrally maintained technical platform, with the capacity to enhance functionality uniformly for all types of users and local management of data since, the research library of the Max Planck Institutes are responsible for the quality control process of the corresponding content.

Perspective: Revolution or evolution?

Presently, the fragmented landscape of institutional repositories predominates, but their repositories are overwhelmingly empty. By contrast, some central publication repositories, not just those supported by funders' mandates but also those built up to serve researchers, have become valuable resources. We therefore contemplate possible transition scenarios, which may help the academic community to go in the direction of more functional repositories with a high value to active researchers.

The first step is to create a community of practice, whereby academic institutions start sharing platforms and deployment strategies. This is not to advocate the deployment of any particular off-the-shelf software, but to suggest that a shared open source project should be established.

The second step is to start sharing developments across technical teams involved in deploying publication repository solutions. This involves two main aspects. First, it minimizes duplication of work by allowing the various partners to focus on specific functionalities and assemble the expertise required to exploit them. Second, it forces the technical team to go deeper into standardizing both their development practices and the actual formats and interfaces they are using. By doing so, one may expect to have an increasing number of generic components at hand.

The third step is more organisational; it has to do with sharing editorial practices between institutions. This ranges from the definition of basic guidelines and priorities that curation should handle uniformly, to the management of joint authority services for institutions, authors, publication places (journals and conferences) and terminology. While sharing practices extensively, this step is also an opportunity to think about

---

[17] For instance, the whole scientific output of a scientist together with some illustrating information. From http://sengbusch.blogs.mpdl.mpg.de/, all publication material are actually referenced from eSciDoc.

[18] The National Institute for Material Sciences (NIMS, Tsukuba, Japan) has adopted eSciDoc as its publication repository platform (NIMS eSciDoc), with the specific aim in mind to offer dedicated researchers' webpages (see http://todoroki.blogs.mpdl.mpg.de/).

[19] E.g. http://colab.mpdl.mpg.de/mediawiki/Faces

specific thematic, sub-institutional or geographic organisation schemes optimizing the editorial support needed for the various communities.

The fourth step builds upon the preceding ones as it corresponds to moving to fully shared deployment platforms across institutions. This is how the target scheme may be achieved, based on coordinated technical work to enhance services at a quick pace and uniformly for all users, as well as a high-level editorial support ensuring the same level of quality requirements for the actual content. This is also the stage at which institutions can actually coordinate their open access policy and contemplate further usages for the publication material as available in the repository.

The specific vision that we have advocated in this paper goes into the direction of providing scientists with digital scholarly workbenches which, through a better coordination of technical infrastructures and adapted editorial support will provide both the quality and flexibility that is required for efficient scientific work. Even if we have focused here on the issue of publication repositories, which, for many reasons, lie currently at the centre of most debates, it is important to consider that this perspective is just one element within a larger set of digital scholarly services that have to be managed in a coordinated way.

Two main directions may be identified. On the one hand, access to subscribed material (online journals, but also eBooks and databases) should not be decoupled from a more integrated repository landscape. Whatever information source is accessed, this should be as seamless as possible for the end-user and all technical decisions (unique identifiers, metadata formats, representation of full-text) should be taken to facilitate this.

On the other hand, attention should be given to the deployment of research data repositories, a complex issue given the variety of type and size of research data, and their integration with publication repositories. A concerted approach could articulate solutions for specific research domains and communities. Existing publication infrastructures may play a stabilizing role by providing coherent concepts that could be seen as global to all scholarly architectures (e.g. attribution and affiliation schemes).

Finally, we should think of coordinating several core services that will increase the effectiveness of scientific digital infrastructures as a whole. Such services may range from cumulative multilingual terminologies to research organisation directories and open unique document identifier schemes. We cannot expect all these to be stable services immediately, but the direction we have tried to follow in this paper aims at showing how a better coordination of scientific infrastructures may allow us to achieve this.

References

Armbruster, C. (2007). Moving out of Oldenburg's Long Shadow: What is the Future for Society Publishing? *Learned Publishing, 20(4),* 259-266. Available at SSRN: http://ssrn.com/abstract=997819

Armbruster, C. (2008a). Cyberscience and the Knowledge-based Economy, Open Access and Trade Publishing: From Contradiction to Compatibility with Nonexclusive Copyright Licensing. *International Journal of Communications Law and Policy, 12*, 22-37. Available at SSRN: http://ssrn.com/abstract=938119

Armbruster, C. (2008b). Open Access in Natural and Social Science: the correspondence of moves to enhance access, inclusion and impact in scholarly communication, *Policy Futures in Education* 6(4) 424-438. Available at SSRN: http://ssrn.com/abstract=849305

Armbruster, C. (2008c). Access, Usage and Citation Metrics: What Function for Digital Libraries and Repositories in Research Evaluation? *Online Currents, 22(5)*, 168-180. Available at SSRN: http://ssrn.com/abstract=1088453

Armbruster, C. (2008d). «Open access» per le scienze sociali. ICavalli, N., Solidoro. A. (eds.) *Oltre il libro elettronico. Il futuro dell'editoria libraria*. Milano: Guerini. Available at SSRN (in English): http://ssrn.com/abstract=846824

Basefsky, S. (2009). The End of Institutional Repositories and the Beginning of Social Academic Research Service: An Enhanced Role for Libraries. Available at LLXR: http://www.llrx.com/authors/1133

Bergstrom, T.C. & Lavaty, R. (2007). How often do economists self-archive? *Department of Economics, UCSB.* http://repositories.cdlib.org/ucsbecon/bergstrom/2007a

Brody T., Gedye R., MacIntyre R., Needham P., Pentz E., Rumsey S. & Shepherd P. (2009). Developing a global standard to enable the recording, reporting and consolidation of online usage statistics for individual journal articles hosted by institutional repositories, publishers and other entities. Final report of project PIRUS – Publisher and Institutional Repository Usage Statistics — http://www.jisc.ac.uk/media/documents/programmes/pals3/pirus_finalreport.pdf

Butler L. (2008). Using a balanced approach to bibliometrics: quantitative performance measures in the Australian Research Quality Framework, *Ethics in Science and Environmental Politics* (ESEP), Vol. 8 - www.int-res.com/articles/esep2008/8/e008p083.pdf

Holmes M. & Romary L. (2009). Encoding models for scholarly literature, in Sarantos Kapidakis (Ed.), *Publishing and Digital Libraries: Legal and Organizational Issues* - http://hal.archives-ouvertes.fr/hal-00390966/fr/

Khayari M., Schneider S., Kramer I. & Romary L. (2006). Unification of multi-lingual scientific terminological resources using the ISO 16642 standard. The TermSciences initiative. In *Proc. of International Workshop Acquiring and representing multilingual, specialized lexicons: the case of biomedicine*, Genoa: Italie - http://hal.archives-ouvertes.fr/hal-00022424/fr/

Poynder R. (2009). Open Access: Whom would you back? Open and Shut? March 10, 2009. - http://poynder.blogspot.com/2009/03/open-access-who-would-you-back.html

Romary L. (2007). OA@MPS - a colourful view, *Zeitschrift für Bibliothekswesen und Bibliographie* - http://hal.archives-ouvertes.fr/hal-00164041/en/

Vogel, B. & Cordes, S. (2005). Bibliotheken an Universitäten und Fachhoschulen. Organisation und Ressourcenplanung. Hannover: HIS Hoschul-Informations-System Gmbh (Hochschulplanung, Band 179) - http://www.his.de/pdf/pub_hp/hp179.pdf