PERSONAL VERSION

Open Access Subject Repositories – an Overview

Bo-Christer Björk
Hanken School of Economics
P.O. Box 479, 00101 Helsinki, Finland
Bo-Christer.Bjork@hanken.fi

Abstract

Subject repositories are open web collections of working papers or manuscript copies of published scholarly articles, specific to particular scientific disciplines. The first repositories emerged already in the early 1990's and in some fields of science they have become an important channel for the dissemination of research results. Using quite strict inclusion criteria 56 subject repositories were identified from a much larger number indexed in two repository indexes. A closer study of these demonstrated a huge variety in sizes, organizational models, functions and topics. When they first started to emerge subject repositories catered to a strong market demand, but the later development of Internet search engines, the rapid growth of institutional repositories and the tightening up of journal publisher OA policies seems to be slowing down their growth.

Introduction

The emergence of the Internet has radically enhanced the possibilities for scientists to disseminate their research ideas and publications directly to potential readers and to bypass the long time delays and selection processes required by traditional publishing. It doesn't cost anything for a scholar to put up a working paper on the web and hope that others will read it, cite it, provide links to it and feedback. Nevertheless the outreach of this is usually not very good, unless the author happens to be a very well-known scientists who's writings are followed by many.

Already before the world wide web paper based services for the dissemination of manuscript stage publications had emerged in certain disciplines and already in 1991 the first Internet-based subject repository, arXiv, emerged (Ginsparg, 2004). Such repositories offer the possibility for authors of embedding their "papers" in a critical mass of other manuscripts on similar topics, which tends to attract more potential readers. It also involves some degree of quality assurance via the brands of the repositories, as well as a more safe and stable storage place compared to personal or departmental web pages. Although subject repositories may also contain article metadata, like traditional citation indexes, as well as research data, most of the interesting ones provide full texts of scholarly publications available free of charge and searchable for web robots.

In a broader context subject repositories provide one of a number of alternative channels for the provision of scholarly Open Access (OA) literature (Suber, 2012). For the peer reviewed journal literature, the primary alternative for OA is that the journals themselves are open access (often called "gold OA"). In 2010 over 8,000 such journals published around 340,000 articles (Laakso and Björk, 2012). This can be achieved by a number of alternative business models, with the model in which authors pay for the publications services rapidly becoming more common (Solomon and Björk, 2012). The other main alternative is author self-archiving of manuscript copies openly on the web (green OA), either on their own or their departments web pages, in the institutional repositories of their universities or in subject repositories, the topic of this study. In an earlier study we found that 20 % of the peer reviewed articles published in 2008 were openly available, with green OA contributing 12 % (Björk et al., 2010). Within green OA subject repositories was the channel for around one third of the articles (Björk et al., 2013).

Literature review

Quite a lot has been written about repositories in general (Armbruster and Romary, 2009, Kim 2010), about author attitudes towards uploading green copies (Nicholas et al., 2012), (Kleinman, 2011) and about the citation advantage of self-archiving in them (Swan 2010, Wagner 2010), but there are few studies that have concentrated specifically on subject repositories. Most of these have presented case studies of individual successful repositories often written by the

scholars who developed them. In their systematic literature review of studies about this topic Adamick and Reznik-Zellen (2010a) in fact state that: “Subject repositories are under-studied and under-represented in library science literature and in the scholarly communication and digital library fields” and further that “The lack of subject repository recognition within the literature … may be attributed to the isolated development of the largest subject repositories and a general lack of awareness about small-scale subject repositories”. In their study they collected papers written after the year 2000 about the ten biggest subject repositories and found only six articles discussing subject repositories more broadly, in contrast to 31 articles discussing individual repositories in rather practical terms.

Kling and McKim (2000) were the first to highlight how the differences in knowledge sharing cultures between scholarly fields prior to the Internet could explain the success of early subject repositories in fields like physics and economics. Darby et al. (2008) looked at the interfaces between subject repositories and institutional repositories, whereas Xia (2008) compared the self-archiving behaviour of physists in both subject and institutional repositories. There are several publications highlighting how individual successful repositories have emerged, in particular the organizational structures that have enabled success (Parinov and Krichel, 2004), (DeRobbio and Katzmayr, 2004), (Ginsparg, 2004), (Kelly and Letnes, 2002). Adamick and Reznik-Zellen also followed up their literature study mentioned above (2010a) with an empirical study of the ten biggest subject repositories (2010b), but little is known about the vast majority of smaller repositories.

The main purpose of this study was thus to get a broader understanding of subject repositories and their development, going beyond the few success stories, in particular looking at the range of varying organizational structures used as well as to get a better understanding of the size distribution, topical range, services, country of origin, and IT platforms used.

Method

There are hundreds of subject repositories included among the more than 2,000 repositories listed in either the Directory of Open Access Repositories (DOAR) or the Registry of Open Access Repositories (ROAR). Also outside the ones indexed in these directories there exists an unknown number of smaller repositories and failed attempts to build repositories.

The only practical way to select repositories for closer scrutiny was to start with the ones indexed in either DOAR or ROAR. Both of these classify repositories into a number of generic types, and it was thus simple to exclude institutional ones (the vast majority) from further investigation. Data about repositories in DOAR listed as either disciplinary (235) or aggregating (96) and in ROAR as research cross-institutional (226) was first collected on the 21 of November 2012. This led to an initial list of 503 candidates, including a lot of duplicates for repositories included in both indexes.

A cursory look at the candidate repositories revealed that a majority of them didn't really fit the description usually given in the OA literature. With this in mind a list of criteria for inclusion in a shorter list were defined.

Repositories which were in scope were such that:

- They have a clear subject limitation
- There must be a channel for authors regardless of affiliation to upload a manuscript as long as it is within the topic area
- At least part of the content consist of working papers and/or submitted or accepted articles
- Access must be open with no charges to most of the publications

Types of repositories, which were discarded included:

- Institutional repositories
- Multi-institution repositories with upload restricted to authors from member institutions only
- Repositories for members of particular associations or projects only
- "Orphan" repositories with no subject limitations (repositories meant for authors from institutions lacking an institutional repository)
- Repositories meant for masters and PhD theses only
- OA Journal portals
- Conference proceedings portals
- Web sites focusing on reusable teaching materials, books
- Historical document archives
- Repositories that charge authors
- Directories with only meta-data
- Portals with just link lists
- Services no longer found with broken links

In some cases it was difficult to draw a line and a couple of border case, but particularly interesting repositories, were included. After the analys 56 repositories remained, ranging from very large repositories with hundreds of thousands of documents to almost unpopulated ones. These were studied using the data available from the ROAR and DOAR indexes, by going to their web sites (in particular the "about" pages) and by searching the web for literature about them. The basic data about these repositories is given in table 1 below.

The chosen 56 are not a random sample of all the repositories first extracted from the indexed repositories. From a method viewpoint this study could instead be labeled a multi-case study. On the other hand the repositories in focus are the ones that fit our own definition of subject repositories oriented towards dissemination of journal articles, and the working papers that precede them.

Table 1. The 56 studied repositories. The topical range has been indicated by the symbol "√" in one of four columns, from left to right: very broad, broad, narrow, very narrow.

Results

Size distribution

The data about the size of the repositories is not particularly accurate, in particular if we are only interested in articles and want to exclude other types of items. For the analysis data from both DOAR and ROAR were used. The data about the number of documents is quite unsecure, the bigger number reported in either ROAR and DOAR has usually been used, and in some cases the number has been checked from the website. In the cases of some repositories it was also possible to directly use the browsing function. For all the reposities using the EPrints software it was possible to extract the exact number of uploaded documents per year.

Despite the potential inaccuraces the data showed that a breakdown into five size categories appeared meaningful. In the following these are shortly presented together with a discussion of typical chararcteristics.

> 100000 documents

In this category we find seven repositories, including: ***PubMed Central (PMC)***, ***CiteSeer***, ***RePEc***, ***arXiv***, ***SSRN*** as well as the less known ***PhilPapers*** and ***Bepress Legal repository***. Four have been started in the 1990's, PMC in 2000, Bepress in 2004 and PhilPapers in 2006. All use custom-built software and have broad topics, covering entire or multiple branches of science, and are located in the US or UK.

> 10000

There are nine repositories in this size category. A couple of them are essentially integral parts of larger repositories (***HAL-SHS*** and ***Munich Personal RePEc*** archive). All use third party software, and all but one have been started after 2000. The majority are located in other countries than the US or UK, particularly in Germany, and also accept content in languages other than English. Some have more specialized topics.

> 1000

This is the most numerous size category (21), and the topics start in many cases to become more narrow. These repositories were with two exceptions founded after 2000 and are spread over several countries. Many have received initial funding from organisations like the European Commission or UNESCO.

> 100

The thirteen repositories in this size range are for the most very narrow in scope, for instance the Basque language (***ArtXiker***) or digital curation and preservation, a subfield of library science (***ERPAePRINTS***). Quite a few use in-house software.

< 100

The five repositories in this size category are largely failed ones, which never reached a critical mass of submissions in order to create a constant flow of new entries. This is very evident in the decline of yearly submissions in later years. It is very likely that there are numerous similar attempts still visible on the web, since such repositories may never have registered in either of the two indexes. Such repositories would be excellent to study in order to find out why they failed, but this has not been attempted here.

Start year

The start years of the different repositories were determined by a number of methods, for instance by checking information on the web site, checking upload data of individual manuscripts and the records in the ROAR registry. The age distribution is shown in Table 2 below, grouped by size category.

Table 2. Repositories by start year, grouped in size categories

Topical range

One useful way of looking at the repositories was to look at how wide or narrow their scope is. The author made a subjective classification of the repositories in four classes: very broad, broad, narrow and very narrow. The yardstick is how roughly how many peer reviewed journal articles are published per year in that topic, and the range was from almost a third of the whole peer reviewed literature in the case of ***PMC***, to a maximum of a few hundred per year for subjects such as ants or IT tools in architectural design. Very broad would correspond to over 100,000 articles per year (ie biomedicine, social science), broad to whole scientific disciplines, often with tens of thousands of articles, narrow to subfields and very narrow to particular topics The results were rather even with 11 very broad, 22 broad, 10 narrow and 13 very narrow.

Services

The basic service that a repository should provide is permanent storage, a stable web adress for the uploaded manuscripts as well as being open to general and web search engines. Most retrievals of the manuscripts would in practice be via search engine hits, for instance using the titles of the articles. Google Scholar has for instance a feature showing openly available full text copies in a separate column to the left.

Many repositories have additional features in addition to this basic functionality. Here are just a few examples:

- Being searchable by special purpose aggregators (ie. OAIster), by complying to the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)
- Possibilty of browsing the repository by subcategories (author, topic, country, year etc, new submissions).
- Advanced search facilty within the repository content
- Citation tracking
- Author ranking based on citations
- Reviews of new manuscripts identifying particularly interesting ones
- Other community services like conference and job announcements

No attempt was made to systematically review such features for each repository. The general impression is that special features could be found especially among the older and bigger repositories which use custom-built software. Most of the younger and smaller repositories, which predominantly have been built using open source repository software, offer the basic set of functionalities.

Organisational setting

The repositories were classified into three types depending on the organisational setting. Those set up by universities, their departements or other organisations (37). Those belonging to international associations (8), and those that have emerged as independent repositories (11), based on initiatives from individual scientists or groups of scientists. Many of the ”independent ” repositories are nevertheless indirectly supported by universities by the free use of their web servers. The difference is mainly in the history and the level of the support by the institution, for instance paid staff managing the repository.

Some of the more succesful independent repositories have over the years evolved substantially in the organisational sense, from their origins of one or a few individual ”entrepreneurs” launching them on their own. ***SSRN*** has in fact become a corporation with a budget in excess of 1 mill USD and an international association has been founded for the running of ***E-LIS***. Some repositories have established hierarchical editorial structures resembling high-quality peer reviewed journals.

It was also interesting to note that several repositories were started with external initial funding, for instance from the Unesco, the European Commission, JISK (UK),US National Science Foundation, Ford Foundation, Nordbib. Most of these were not particularly successful in attracting a steady flow of new submissions after the intial funding ended.

IT platform

All the early repositories had to build IT platforms of their own, but later most repositories have been able to use specialised third party software (ie EPrints and DSpace), usually available via Open Source licenses. Seventeen repositories use custom built software, in particular the biggest and most successful ones. EPrints (17 repositories) and DSpace (10 repositories), which initially were designed for institutional repositories, are both very popular platforms for smaller and mid-sized subject repositories. DLIST (Digital Library of Information Science & Technology) has in fact been implemented as a subject collection in the University of Arizona institutional DSpace repository. HAL (Hyperarticles en ligne) is the French national repository infrastructure which also has been used to set up a number of subject collections, of which six are included in this analysis. Opus with four repositories is a system developed by the University of Stuttgart and is widely used in Germanspeaking countries and Fedora (one repository) was originally developed by researchers at Cornell University. SciX was developed in an EU funded project by Ljubljana University and is used in one repository.

Country of origin

Not unexpectedly, the US topped the list with 17 repositories. More surprisingly Germany tied with the UK in second place with 9 repositories each. The German based repositories where for the most rather young, many using the OPUS software and also containing a fair proportion of German language content (in addition there were two Austrian repositories). France had 4 and Italy 3, with all other countries having at most two.

Repositories by discipline

There are many alternative ways in which a discussion of the 56 repositories could be organized. The most meaningful seems to be by subject field. The narrative is also partly chronological, in the sense that the fields where subject repositories first developed start off the discussion. Table 1 above has been structured to follow more or less the same order as the narrative.

Physics and mathematics

Scientists in all disciplines tend to send article manuscripts to a few colleagues to get feedback and exchange ideas but in a few fields such as physics and economics this exchange was more systematic even before the Internet and the Web, first on paper and later using ftp sites and email list servers. It was thus no coincidence that the first successful repositories emerged in these subject fileds. Much has been written about the first successful E-print archive, ***arXiv***, which was started in 1991 by Paul Ginsparg at the Los Alamos National Laboratory. The number of manuscripts uploaded to arXiv has in two decades grown linearily to over 800,000, and nowadays the repository includes other disciplines, such as mathematics, non-linear science, computer science, and quantitative biology. In 2001 Cornell University took over the hosting of the service, and in 2011 university paid staff have taken over the practical running of the repository. Its

current yearly budget is around 400,000 USD and the university is trying to get institutions with many authors uploading to the repository to support the service financially. A collaborative governance structure has also been set up for the repository (Fischman, 2011).

Despite the success of arXiv, a number of scientists who are dissatisfied with certain aspects of its operation, have set up an alternative service called ***viXra***. According to the latter's website "It has been founded by scientists who find they are unable to submit their articles to arXiv.org because of Cornell University's policy of endorsements and moderation designed to filter out e-prints that they consider inappropriate". The service, founded in 2009, currently houses over 3,000 manuscripts.

Economics and Management

The other branch of science, which has had a strong preprint culture preceding the web is economics, where manuscripts have been distributed as working papers published by the university departments and research institutes of the authors. The ***Social Sciences Research Network (SSRN)*** was started in 1992 under the name Financial Economics Network and was formally incorporated under its current name in 1994. It has a budget of currently around 1 mill USD but is largely dependant on voluntary work contributed by over 1,000 scholars worldwide, who act as Advisory Editors, Editors and Network Directors (Jensen 2012). The paid staff includes about 15 people in the central office.

SSRN currently stores over 380, 000 full papers and abstracts, and while the majority of papers are available for free the network includes a number of indexing and abstracting services which are subscription based. SSRN also includes material from publishers who upload it to SSRN, but which material can only be accessed via pay-per-view.

RePEc (Research papers in economics) was started in 1997, but its origins strech back to the beginning in 1993 when its predecessor NetEc was started. In contrast to SSRN, RePEc is entirely run by volunteers and all services are free. The structure of RePEc differs from most subject repositories since its backbone consists of a large number (currently over 1,400) of institutional or departmental repositories of working papers and preprints, which are linked together by a common search portal and a number of value-adding services, for instance download statistics. Author's who lack a suitable local repository to upload to, can use the ***Munich Personal RePEc Archive***. There is also an adaption of the same software and structure in Russian; ***Socionet***. In addition major publishers provide metadata info of published subscription articles.

Both SSRN and RePEc are very representative examples of the web portal philosophy, which was very popular in the latter half of the 1990's, in the period where web search engines were not yet fully developed, and when readers interested in particular subjects would tend to search for information in discipline-specific hubs.

EconStor is a much more recent initiative (2009), and also has a different type of genesis. It's predecessor was the German National Library for Economics and it

migrated to the DSpace software in 2009. It has a very clear institutional setting and provides repository services to the economics departments of German universities. It's development has partly been supported by the European Commission (the NEEO project). EconStor also contains interfaces to RePEc. EconStor is less of a portal where researchers would directly search for publications than a method for a large number of institutions to outsource the technical infrastructure for their repositories.

Computing and Information Science

It is perhaps no big surprise that researchers in computing and information&library science, given their research areas, have been very active in the creation of subject repositories. ***CiteSeer*** was started in 1997 by a number of researchers working at the NEC research institute. It was not primarily a repository but rather a search engine for academic content, which harvests the web for openly available literature, and also provides citation tracking functions. It can be seen as forerunner to academic search engines such as Google Scholar. In 2008 it was replaced by ***CiteSeerx***, which has a more scalable software architecture.

E-LIS, e-prints in library and information service, was established in 2003 by an international group of collaborating scholars, and has organizationally evolved to an association, with a structure similar to top-notch society published scholarly journals. The organization includes nearly 50 national editors, which check the meta-data of documents uploaded from their respective countries.

Other repositories in computing and information science have mainly been limited to output from particular countries (the French ***Archivesic*** and the ***Arab Repository for Library and Information Studies***) or to narrow subject fields. Examples of such fields are agent-based computing, cryptology, graph drawing, information systems, digital curation & preservation. This author has personal experience of two such repositories, ***Sprouts*** for working papers in Information systems and ***Architektur-Informatik*** for IT in architecture. Both were created with great enthusiasm just after the millennium shift but have never achieved the critical mass of submissions needed for success.

Medicine

The medical field has only few repositories, but one, ***PubMed Central (PMC),*** is a key resource for the whole open access movement. PMC was developed by the US National Library of Medicine based on the earlier Entrez search engine for health sciences databases and launched in 2000. In contrast to many of the repositories mentioned earlier PMC is concentrated on providing open access to manuscript copies of published articles, not working papers or submitted versions. Many publishers have also agreed to deposit exact copies of published papers in PMC, usually with a delay of 12 months. Currently PMC contains over 2 million open access articles.

Of particular importance for the development of PMC has been the OA mandate of the National Institutes of Health, which has been in place since 2006. This policy requires that articles emanating from NIH funded research are made

openly available in PMC at the latest 12 months after publication. Due to NIH's position as the biggest public research funding agency in the world, a number of publishers have lobbied strongly against this mandate in the US legislation, however unsuccessfully. Due to the great popularity of PMC, sister sites have sprung up in other countries (***PMC Canada***, ***UK PubMed Central***)

One of the reasons there have been fever repositories founded in biomedicine is the absence of a working paper or preprint tradition in this field, and the relative fast turnaround time from submissions to published articles in medical journals. One attempt to set up a preprint server in the field was ***Clinical Medicine NetPrints***, which was sponsored by the BMJ group and Highwire Press. The site contains less than a hundred manuscripts from 1999 to 2003. The home page of the repository has a very visible warning: "Articles posted on this site have not yet been accepted for publication by a peer reviewed journal. They are presented here mainly for the benefit of fellow researchers. Casual readers should not act on their findings, and journalists should be wary of reporting them".

There are a few other repositories in medicine worth mentioning. ***OpenMED@NIC*** is hosted by the Bibliographic Informatics Division of National Informatics Centre (India) and is intended for both preprints and green copies of accepted manuscripts. It seems to be mainly used by Indian authors. ***Dryad*** offers authors of published medical articles the possibility to upload data sets related to their journal articles to the repository.

Philosophy

Scientists in Philosophy seem to have eagerly embraced the idea of subject repositories. ***PhilPapers*** also includes other types of materials than just OA copies of articles, and has many characteristics of a "one-stop shop" portal for scientists in the domain. Like many other subject repositories it has been developed (since 2006) by a couple of entrepreneurial scientists, although it has received sponsorhip from JISC in the UK. ***Sammelpunkt. Elektronisch Archivierte Theori*** is a small repository with a wide variety of subjects, although the majority of papers are in some field of philosophy. It contains papers in both German and English.

There are two repositories with more narrow subject areas worth mentioning. The ***Philosophy of Science*** archive was set up in the year 2000 by scholars at the University of Pittsburgh, inspired by the success of arXiv. Like its role model it is concentrating on preprints. ***SciRePrints*** is a small repository hosted by the University of Latvia, focusing on papers discussing the relationship between science and religion. The home page contains an interesting passage: "Notably, scientific articles not accepted for publishing in other sources due to religious or mystical presuppositions are welcome here, provided that they comply with the academic standards and use scholarly methods and language".

Earth Sciences

There are several repositories within the general area titled Earth Sciences, including two broad one dealing with earth and atmospheric sciences (***CEDA***, ***Earth Prints***) and more specialized one dealing with topics like Marine

environment research (***Aquatic Commons***), Organic Food and farming (***Organic Eprints)*** and research related to the Carribean Region. A successful repository with a critical mass of material is ***AgEcon Search***, hosted by the university of Minnesota, which specializes in agricultural and applied economics (Kelly and Letnes, 2006). Another narrowly focused one is ***Antsbase***, a collaboration between the American Museum of Natural History and the Ohio State University.

Social Sciences

In addition to Economics and Management there are several repositories in the social sciences, but only three with a significant critical mass of papers exceeding 10000 manuscripts (***Social Science Open Access Repository***, the German ***eDOC.VifaPol*** for administrative and political science and ***HAL-SHS***). The last one of these is one of the overlay structures providing subject views, in this case social sciences, into the national French HAL repository. In addition there are several more specialized ones, for instance five in education and pedagogy and three in psychology and psychiatry. An interesting one combining a subject and regional aspect is ***African Higher Education Research Online***.

In history research there were several repositories listed in ROAR and DOAR which had to be discarded because they contained digitized documents only, but there were interesting repositories for instance dealing with European Integration and Latin American development. A very well functioning repository seems to be the ***Forced Migration Online Digital Library***, which aims to collect a multitude of information resources related to refugees and forced migration (Cave et al., 2008). The web site is of a high quality and the repository aims to also reach out to policy makers, the broader public and teachers. The web site also provides a facility to donate money for the maintenance of the site.

Arts and humanities

There are several highly specialized repositories in the arts and humanities, but no broader ones. It is important to note that scholars in these fields tend to publish more in monographs or book chapters, and that peer reviewed journal publishing is less compared compared to the STM sciences.

Curiously there are two repositories dealing with different aspects of Basque culture, one more broader (***Hedatuz***) and one concentrated on the language (***ArtXiker***) Both accept inputs in several languages (Basque, French, Spanish, English). The latter, like ***hprints.org*** (The free Nordic Arts and Humanities and Social Sciences e-print repository) uses the French national repository infrastructure HAL. Hprints.org was started with Nordic funding but never really took off. Other repositories with narrow domains exist for classical studies (***Propylaeum-DOK***), art history (***ART-Dok***) and archeology (***JIIA***).

Discussion

The evolution of subject repositories must be seen in context against the general development of the Internet and the development of the open access movement.

Several of the leading repositories were developed already in the mid 1990's when most scholarly journals were still only distributed as paper copies, and when creating portals to web information and link lists was more important than today. Since then almost all major publishers have started parallel electronic publishing of subscription journals and a rapidly increasing number of universities and research institutes have launched institutional repositories of their own, which compete with subject repositories for the same papers.

Only the biggest subject repositories contribute significantly to the overall volume of green OA copies. A recent study found that 43 % of self-archived manuscript copies are located in subject repositories (Björk et al., 2010). 94 % of these were located in either arXiv or PMC (Björk et al., 2013). Most of the other repositories may play an important role in their niche areas, but there are so many blank research areas without a relevant subject repository that the overall effect is very small. Areas, which in view of this study lack significant repositories, are for instance chemistry and engineering.

Determining if a particular repository is successful is a difficult and subjective task. On a theoretical level methods for measuring this for any type of repository are for instance discussed by Thibodeau (2007) who states that success is measured by " how well it covers the universe of assets it should or might hold". Also (Adamick, J. & Reznik-Zellen, R. 2010b) discuss this issue. The success or critical mass of a repository should ideally be judged by comparing the actual uploaded content to the potential uploadable literature in the topical range. In practice it would, however, be very difficult to determine the potential article volumes for many of the repositories, unless their topics coincide exactly with disciplines defined in Web of Science or Scopus. Equally it would be difficult to find out the exact numbers of WoS or Scopus indexed articles among the documents uploaded to the repositories, since they many contain a wide variety of material (alos other than copies of peer reviewed articles). A pragmatic measure which is easier to use, is the trend in the number of uploads, which has also been discussed by Carr and Brody (2007). Authors in a field will soon loose faith in a repository, which hasn't achieved a critical mass of article and will stop uploading new documents. For some of the smaller repositories, which in the discussion section have been mentioned as failures, this criterion has been used.

It is interesting to note the wide variety of organizational structures, which have emerged around subject repositories. The most common history is that of a single or a handful of "entrepreneur" scholars who have created the repository as a more or less personal project. Usually their institution has allowed the use of the university website. In some cases the development has later lead to a corporate structure with employed staff (SSRN), in others to the emergence of complex networked voluntary work structures (RePEc). Repositories, which have been started by institutions on a strategic decision by top management (i.e. PMC, HAL) are rare. Armbruster and Romary (2009) note that "the future of subject-based repositories depends on whether they develop a sustainable business model with independent income".

The results of this study can be compared with (Adamick, J. & Reznik-Zellen, R. 2010b), who studied the ten biggest repositories also included here. The following observations can be made.

- The geographical spread of the home countries of the repositories becomes much more diverse (outside the US) as we go outside the "big 10". Several midsized and small repositories also welcome uploads in other languages than English.
- Midsized and smaller repositories tend to have more narrow topical ranges than the big ones
- There are several smaller repositories in niche areas in the social sciences, arts and humanities.
- The use of open source software is dominant outside the big 10.
- There are many failed repositories included. The business model where a repository was started with time-limited funding from an outside grant seems not to have been particularly fruitful.

Conclusions

Despite the availability of open source repository software (i.e. EPrints, DSpace), which technically has lowered the initial effort needed to start a repository, it seems that the potential for launching new successful repositories has diminished. It is currently much from managerial viewpoint easier to launch an institutional repository, which is usually operated by dedicated university library staff, than a subject repository, which may lack initial funding and requires an international network of collaborators to get going. An IR is the natural locus of Ph.D. and lower theses, already existing working paper and publication series, and it can be backed by a mandate from the university making it obligatory for the institutions researchers to upload green copies of their journal publications. Institutional repositories are also natural extensions to the Current Research Information Systems, that almost all universities now have to keep track of their publication output, and it is very "trendy" to start IRs. According to Björk et al (2013) 82 % of the world's 148 most productive research institutions have an institutional repository offering a natural place to self-archive for 85 % of the articles produced in these institutions. Subject repositories, on the other hand, must rely mainly on word-of-mouth within their communities and on reaching the necessary critical mass early in order to take off.

Another development of importance is in the legal boundary conditions for self-archiving. Although a majority of publishers allow the open self-archiving of the final approved manuscript version, this is usually allowed only for author web pages and institutional repositories, usually subject repositories are excluded. In a study of the copyright rules of the 100 largest scholarly publishers, immediate self-archival of the accepted version was allowed for 61 % of all published articles in institutional repositories but for only 21 % in subject repositories (Laakso, 2013). Such detailed copyright rules started to become common around 2003-2004, and apparently publishers feel that subject repositories are a bigger threat to their business. The only exception is PMC, which due to the bargaining power of NIH as research funder, has been able to negotiate special conditions with several major publishers.

The few big successful subject repositories are likely to continue to thrive, because they have become a part of the publishing behavior of academics in their fields. Two factors have in particular contributed strongly to the emergence of successful subject repositories in a limited number of areas. Firstly the existence of a strong working paper or preprint culture in a research field already prior to the Internet (as was the case for arXiv, SSRN and REPEC) and secondly a mandate from a dominating research funder to upload copies to a prescribed subject repository (PubMed Central). If new mandates from strong research funders would emerge (for instance through stakeholders such as the US government or the European commission) these might help support the growth of existing or new subject repositories, but the trend in such mandates seems to be to promote self-archiving in institutional repositories as well, which is not the case in the NIH mandate.

All in all it seems that the strongest growth period for subject repositories is over. The growth in green OA literature available via subject repositories currently mainly consists of the internal growth of the few really big ones (PMC and arXiv in particular) rather than emergence of new repositories.

Acknowledgements:

The author wishes to thank the two anonymous reviewers for very constructive comments, which have helped a lot in improving the manuscript.

References:

Adamick, J. & Reznik-Zellen, R. (2010a). Representation and Recognition of Subject Repositories, D-Lib Magazine, 16, doi:10.1045/september2010-adamick

Adamick, J. & Reznik-Zellen, R. (2010b). Trends in Large-Scale Subject Repositories, D-Lib Magazine, 16, doi:10.1045/november2010-adamick

Armbruster, C. & Romary, L. (2009). Comparing repository types: Challenges and barriers for subject-based repositories, research repositories, national repository systems and institutional repositories in serving scholarly communication, Working paper, November 23, 2009. Retrieved from: http://ssrn.com/abstract=1506905.

Björk, B-C., Laakso, M., Welling, P. & Paetau, P., (2013). Anatomy of Green Open Access, In press, Journal of the American Society for Information Science and Technology.

Björk, B-C., Welling, P., Laakso, M., Majlender, P., Hedlund, T. & Gudnasson, G. (2010). Open Access to the Scientific Journal Literature: Situation 2009 PLoSOne, 23.6.2010, doi:10.1371/journal.pone.0011273

Björk, B-C. & Solomon, D. (2012). Open access versus subscription journals: a comparison of scientific impact, BMC Medicine, 10:73, doi:10.1186/1741-7015-10-73

Carr, T. & Brody, T. (2007). Size Isn't Everything – Sustainable Repositories as Evidenced by Sustainable Deposit Profiles, D-Lib Magazine, 13, http://www.dlib.org/dlib/july07/carr/07carr.html

Cave, M., Loughna, S. & Pilbeam, J. (2008). Open Access Repository System for Forced Migration Online, Association of Librarians and Information Professionals Quarterly, 3(4)

Darby, R., Jones, C., Gilbert, L., & Lambert, S. (2008). Increasing the productivity of interactions between subject and institutional repositories, New Review of Information Networking, 14, 117-135.

DeRobbio, A., & Katzmayr, M. (2009). The management of an international open access repository: the case of E-LIS.GMS Medizin - Bibliothek - Information, 9(1), http://www.egms.de/static/pdf/journals/mbi/2009-9/mbi000137.pdf

Fischman, Josh (2011) The First Free Research-Sharing Site, arXiv, Turns 20 With an Uncertain Future, The Chronicle of Higher Education, 10.8.2011, Retrieved from: http://chronicle.com/blogs/wiredcampus/the-first-free-research-sharing-site-arxiv-turns-20/32778?sid=wc&utm_source=wc&utm_medium=en

Ginsparg, P.(2004). Scholarly Information Architecture, 1989-2015. Data Science Journal, 3, 29-41. Retrieved from: https://www.jstage.jst.go.jp/article/dsj/3/0/3_0_29/_pdf

Jensen, M. (2012) About SSRN, 2.2.2012, Retrieved from: http://www.ssrn.com/update/general/mjensen-20th.html

Kelly, J., & Letnes, L. (2006). Managing the grey literature of a discipline through collaboration: AgEcon search. Resource Sharing & Information Networks, 18, 157-166.

Kim, J. (2010). Faculty self-archiving: Motivations and barriers, Journal of the American Society for Information Science and Technology, 61, 1909–1922.

Kleinman, M. (2011). Faculty self-archiving attitudes and behavior at research universities - a literature review, Ph.D. term paper, University of Michigan, Retrieved from:

http://mollykleinman.com/wp-content/uploads/2012/02/Kleinman-self-archiving-literature-review-web.pdf

Kling, R. & McKim , G. (2000). Not Just a Matter of Time: Field Differences and the Shaping of Electronic Media in Supporting Scientific Communication, Journal of the American Society for Information Science and Technology, 51, 1306–1320.

Laakso, M. (2013). Journal publisher self-archiving policies and the potential for growth in open access. Working paper, Hanken School of Economics: Helsinki, Finland.

Laakso, M. & Björk, B-C. (2012). Anatomy of open access publishing - a study of longitudinal development and internal structure, BMC Medicine, 10:124, doi: 10.1186/1741-7015-10-124.

Nicholas, D., Rowlands, I., Watkinson, A., Brown, D., & Jamali, H. R. (2012). Digital repositories ten years on: what do scientific researchers think of them and how do they use them?, Learned Publishing, 25, 195–206.

Parinov, S. & Krichel, T. (2004). RePEc and Socionet as partners in a changing digital library environment, 1997 to 2004 and beyond. In: Russian Conference on Digital Libraries, Puschchino, Russia. Retrieved from: http://eprints.rclis.org/1830/.

Suber, P. (2012). Open Access. Boston: MIT press, http://cyber.law.harvard.edu/hoap/Open_Access_(the_book)

Swan, A. (2010). The Open Access citation advantage: Studies and results to date. Key Perspectives Report. Retrieved from: http://eprints.ecs.soton.ac.uk/18516

Thibodeau, K. (2007) If you build it, will it fly? Criteria for success in a digital repository, Texas Digital Library, 8, http://journals.tdl.org/jodi/index.php/jodi/article/view/197/174

Wagner, B. (2010). Open Access Citation Advantage: An Annotated Bibliography. Issues in Science and Technology Librarianship. doi:10.5062/F4Q81B0W

Xia, J. (2008). A comparison of subject and institutional repositories in self-archiving practices. Journal of Academic Librarianship, 34, 489-95.

	Range				Items	Founded	Country	Software	Type
PHYSICS AND MATHEMATICS									
arXiv	√				805000	1991	USA	In-house	Independent
viXra	√				3680	2009	UK	In-house	Independent
ECONOMICS AND MANAGEMENT									
Social Science Research Network	√				814725	1992	USA	In-house	Independent
Research Papers in Economics		√			1200000	1993	USA	In-house	Independent
Munich Personal RePEc Archive		√			22643	2006	Germany	EPrints	Institution
Socionet		√			3520	2006	Russia	In-house	Independent
Econstor		√			48252	2009	Germany	DSpace	Institution
Industry Studies Working Papers				√	130	2010	USA	EPrints	Association
COMPUTING AND INFORMATION SCIENCE									
CiteSeer^{X}	√				2000000	1997	USA	In-house	Institution
E-LIS		√			14053	2003	Italy	DSpace	Association
Arab Repository for Libr. and Inf. Studies		√			52	2010	Egypt	In-house	Institution
Archivesic		√			1501	2002	France	HAL	Institution
DLIST		√			1540	2002	USA	DSpace	Institution
Sprouts				√	485	2000	USA	In-house	Association
Cryptology ePrint Archive				√	5702	1996	USA	In-house	Association
Architektur-Informatik				√	113	2003	Austria	SciX	Independent
ERPAePRINTS				√	82	2003	UK	EPrints	Institution
AgentLink				√	1410	2004	UK	EPrints	Association
Graph Drawing E-print Archive				√	886	2003	Germany	EPrints	Institution
MEDICINE									
PubMed Central	√				2600000	2000	USA	In-house	Institution
OpenMED@NIC	√				2866	2005	India	In-house	Institution
Clinical Medicine NetPrints	√				81	1999	UK	In-house	Association
Dryad	√				8849	2009	USA	DSpace	Institution
PHILOSOPHY									
PhilPapers		√			507277	2006	UK	In-house	Institution
PhilSci Archive			√		3005	2000	USA	EPrints	Institution
Sammelpunkt. Elektronisch archivierte Theorie	√				1526	2002	Austria	EPrints	Independent
SciRePrints				√	164	2009	Latvia	EPrints	Institution
EARTH SCIENCES									
CEDA Repository		√			812	2009	UK	EPrints	Institution
Earth-prints Repository		√			7780	2006	Italy	DSpace	Institution
Aquatic Commons			√		8072	2007	Belgium	EPrints	Association
Organic Eprints				√	13013	2002	Denmark	EPrints	Institution
AgEcon Search			√		58007	1995	USA	DSpace	Institution
Open Knowledge Environment of the Caribbean			√		8	2009	Jamaika	DSpace	Institution
antbase.org				√	500	1995	USA	In-house	Institution
SOCIAL SCIENCES									
Social Science Open Access Repository	√				21777	2007	Germany	DSpace	Institution
HAL-SHS	√				41416	2003	France	HAL	Institution
eDoc.VifaPol		√			66070	2000	Germany	Opus	Institution
Bepress Legal Repository		√			134931	2004	USA	In-house	Institution
EduDoc		√			488	2008	Mexico	In-house	Institution
Fachlicher Dokumentenserver Paedagogik		√			4414	2005	Germany	In-house	Institution
Cognitive Sciences ePrint Archive		√			4010	1997	UK	EPrints	Independent
African Higher Education Research Online			√		828	2007	S.Africa	In-house	Institution
Kaleidoscope Open Archive				√	1357	2006	France	HAL	Association
PsyDok		√			2319	2004	Germany	Opus	Institution
Theory of Psychology Eprint Archive			√		119	2001	UK	EPrints	Independent
Bibliopsiquis		√			4789	2001	Spain	DSpace	Association
Policy Archive			√		21935	2008	USA	DSpace	Institution
Archive of European Integration				√	20280	2003	USA	EPrints	Institution
Latin American Development Archive			√		12	2007	USA	EPrints	Institution
Forced Migration Online Digital Library				√	4827	2002	UK	Fedora	Institution
ARTS AND HUMANITIES									
Hedatuz			√		8133	2002	Spain	EPrints	Institution
ArtXiker				√	394	2006	France	HAL	Institution
Hprints		√			116	2008	Denmark	HAL	Institution
ART-Dok		√			2551	2007	Germany	Opus	Institution
Propylaeum-DOK		√			1536	2007	Germany	Opus	Institution
JIIA Eprints Repository		√			200	2003	Italy	EPrints	Independent

Table 1. The 56 studied repositories. The topical range has been indicated by the symbol "√" in one of four columns, from left to right: very broad, broad, narrow, very narrow.

Start year	≥100000	≥10000	≥1000	≥100	≥1	sum
1991	1					1
1992	1					1
1993						0
1994						0
1995			1		1	2
1996				1		1
1997	2			1		3
1998						0
1999					1	1
2000	1		1	1	1	4
2001				1	1	2
2002			1	5		6
2003			3		3	1	7
2004	1			2		3
2005				2		2
2006	1		1	3	1	6
2007			1	4	1	1	7
2008			1		3		4
2009			1	1	1	1	4
2010					1	1	2
2011							0
2012							0
SUM	6	9	21	13	5	56

Table 2. Repositories by start year, grouped in size categories