informationresearch

PUBLISHED QUARTERLY BY THE UNIVERSITY OF BORÅS, SWEDEN

VOL. 22 NO. 2, JUNE, 2017

Institutional repositories as infrastructures for long-term preservation

Helena Francke, Jonas Gamalielsson, and Björn Lundell.

Introduction. *The study describes the conditions for long-term preservation of the content of the institutional repositories of Swedish higher education institutions based on an investigation of how deposited files are managed with regards to file format and how representatives of the repositories describe the functions of the repositories.*
Method. *The findings are based on answers to a questionnaire completed by thirty-four institutional repository representatives (97% response rate).*
Analysis. *Questionnaire answers were analysed through descriptive statistics and qualitative coding. The concept of information infrastructures was used to analytically discuss repository work.*
Results. *Visibility and access to content were considered to be the most important functions of the repositories, but long-term preservation was also considered important for publications and student theses. Whereas a majority of repositories had some form of guidelines for which file formats were accepted, very few considered whether or not file formats constitute open standards. This can have consequences for the long-term sustainability and access of the content deposited in the repositories.*
Conclusion. *The study contributes to the discussion about the sustainability of research publications and data in the repositories by pointing to the potential difficulties involved for long-term preservation and access when there is little focus on and awareness of open file formats.*

Introduction

Over the past twenty years, significant efforts have been made to provide open access to research publications, and more recently also to research data. Making publications available through open access is something that is increasingly mandated by funders and universities in many countries. The open access movement has led to new business models through which publications are made open access by the publisher (gold open access), but also to the set-up of repositories for parallel publishing of scholarly work (green open access) in order to assist authors in making versions of their published, toll-access, work accessible to readers more broadly and, in some cases, available for automatic indexing and mining. Furthermore, institutional repositories are often used for gold open-access publishing of theses and dissertations, grey literature such as working papers and reports and occasionally research data. In this article, we focus on institutional repositories. Their work is often perceived as making the publications of a higher education institution or research institute available to a wide audience without charge and as soon as possible after publication, although the time between primary and parallel publication is often regulated through publisher-initiated, copyright-motivated, embargo periods. We investigate how institutional repository management view the repositories' functions in a long-term perspective and how sustainable is the content in the repositories.

It is not self-evident that institutional repositories should have a long-term and future-oriented perspective. Many of the documents deposited in the repository, which have their primary site of publication elsewhere, can also be expected to be part of other long-term preservation plans. Furthermore, the repository often does not contain the publisher's version of the work. However, there are also reasons why institutional repositories may wish to plan with the future in mind. To begin with, some publications in the repositories may not be published elsewhere. This is often the case with student theses and doctoral dissertations, as with report series, digitised material and open educational resources (Ball, 2010). It may also be of relevance to the institution to maintain copies of their employees' publications in a central database and in forms that may be used in various ways.

Inspired by a socio-technically oriented view of information infrastructures (e.g., Bowker *et al.* , 2010), we view institutional repositories and their development as constructed within a nexus of social, political, technological, organizational, economical and ethical choices. Here, we will be concerned with choices made from primarily two perspectives. First, at the organizational level, the institution's plans and ambitions for the repository are central to policy work as well as to the day-to-day management of the repository. It is therefore relevant to investigate the organization's views on the repository's functions and its life-span.

Secondly, in the management of the repository, choices made with regard to technology will influence the future accessibility of the repository. The use of open and well-documented file formats (and open source software) is one way to increase the possibility of accessing, migrating and in various ways extracting information from publications in the future. Therefore, this study addresses questions concerning what file formats are accepted and stored in the repositories and how the files are managed. Other issues of great importance for long-term preservation concern, for instance, metadata about rights management, and about which versions of publications are stored. However, we have chosen to limit our focus to technological features in this paper.

The aim of our study is to describe the conditions for long-term preservation of the content of the institutional repositories of Swedish higher education institutions and thus contribute to the discussion about the sustainability of research publications and data in the repositories. This will be achieved by addressing the following research questions:

- How do Swedish institutional repositories restrict, promote, manage and document file formats of files deposited in the repositories?
- How do respondents at Swedish institutional repositories describe the functions of the repositories?
- What are some expected consequences of the answers to these two questions for the content stewardship in the institutional repositories to ensure the sustainability of the content?

The research is based on a survey sent to higher education institution repositories in Sweden in the summer of 2015.

Institutional repositories in Sweden

Swedish higher education institutions are of various sizes and levels of specialisation, and they have varying organizational structures. The Swedish Higher Education Authority (2015) makes a distinction between higher education institutions that have the right to award qualifications up to and including the third-cycle level (doctoral level, twenty-eight higher education institutions), and those that may award first and second-cycle qualifications (twenty institutions). Of the former group, all institutions have institutional repositories. Furthermore, a distinction is made between the thirty-one institutions that are accountable to the government, thirteen independent institutions, and four independent course providers. All institutions of the first type, with two exceptions, have institutional repositories, as do six of the independent institutions (three institutions that offer third-cycle qualifications and three other institutions, all within nursing). The institutions that lack a repository are all small and highly specialised within a particular area of education, often in the fine arts, theology or psychotherapy.

At the time of the survey, and as stated in the responses to the survey, the majority of the institutions (twenty-seven) used DiVA, a repository platform developed in Sweden and maintained through a membership consortium. This dominance of DiVA means that the situation diverges from that in other countries (Chowdhury, 2014). Five institutions used more than one system. This was in some cases because they were transitioning from one system to another and in other cases because different systems were used for different purposes. Other systems in use were DSpace (4), Converis (2), Scigloo, developed by Chalmers University of Technology and the University of Gothenburg (2), EPrints (1), Pure (1), Librecat, co-developed by the University of Lund (1), Lotus (1) and S-WoBA/S-WoPEc, for subject repositories (1). It should be noted, however, that platform changes have been taking place since summer 2015 and that the figures describe the situation at the time of the data collection. For most of the institutions, the platform is used both as a publication database (listing metadata for all employees' publications) and as an institutional repository (with uploaded files). The widespread use of DiVA means that the platform use is quite homogeneous. All institutions also upload their metadata to the national publication database SwePub, which provides a joint interface to search for research publications from Swedish higher education institutions.

Related research

A substantial body of research has, over the past decade, been devoted to open access to scientific and scholarly publications, and recently also to research data. Similarly, there is a fair amount of research aimed at understanding the life cycle of digital files and the openness of file formats and software. However, little research has so far been devoted to the intersection of these two research areas in the context of academic publishing (Termens, Ribera and Locher, 2015; Sawant, 2011; and Rimkus, Padilla, Popp and Martin, 2014 are three exceptions). Broader discussions of repository sustainability (e.g., Rieger, 2012; Eschenfelder and Shankar, 2016) as well as on best practice and tools in preservation and curation (e.g., Ball, 2010; Robertson and Borchert, 2014) mention the importance of standards compliance but do not go into details and do not elaborate on the important difference between a file format as documented in a technical specification and a file format as implemented in software.

Institutional repositories

Since the early 2000s, institutional repositories have become a common infrastructure in higher education institutions world-wide. In investigating if an institutional repository was available for authors, Björk, Lakso, Weiling and Paetau (2014) studied the 148 top universities in terms of publications in Scopus and concluded that 82% of them had at least one institutional repository. OpenDOAR, a directory of open access repositories, lists 2,578 institutional repositories spread across 117 countries, as of February 2016 (OpenDOAR, 2014). Even so, the availability of institutional repositories is unevenly distributed, with the majority located in North America, Europe and Japan (Chowdhury, 2014, p. 122; Cho, 2014).

Stevenson and Zhang (2015) showed that research on information repositories has matured as a field, especially since 2010, and that much of the research comes from library and information science and in the form of case studies (see also Björk *et al.*, 2014). Cho's (2014) co-word analysis of the institutional repository research area showed that much research has concerned metadata and interoperability. The study identified preservation as a related but separate domain.

In describing the function and mission of institutional repositories, several authors draw on definitions made by Clifford Lynch, such as the following, which portrays institutional repositories as (Lynch, 2003, p. 328):

> *a set of services that a university offers to the members of its community for the management and dissemination of digital materials created by the institution and its community members. It is most essentially an organizational commitment to the stewardship of these digital materials, including long-term preservation where appropriate, as well as organization and access or distribution.*

In this early description, Lynch suggests that long-term preservation is a key function for the repository, along with access and distribution. This focus on storage is not surprising, given that the databases are referred to as *repositories*. However, the databases have in many cases come to serve several functions, not least as databases of publication metadata used for bibliometric analysis of research output. In Sweden, the perceived need for monitoring research output and providing open access to publications have been parallel motivations for the implementation of publication databases, which in almost all cases also function as institutional repositories (Francke, 2013; see also, Research Information

Network. *Working Group...*, 2012).

Functions of institutional repositories

Motivations for setting up an institutional repository, and the functions they serve, are diverse. Access to documents as well as their preservation, both mentioned by Lynch (2003) in the quotation above, are functions commonly dealt with in the literature (e.g., Kunda and Anderson-Wilk, 2011; Sawant, 2011; Jacobs, Thomas and McGregor, 2008; Termens, Ribera and Locher, 2015; Research Information Network. *Working Group...*, 2012; Rieh, St Jean, Yakel, Markey and Kim, 2008; Kennan and Wilson, 2006; Probets and Jenkins, 2006; Ball, 2010; Robertson and Borchert, 2014; Jones, Darby, Gilbert and Lambert, 2008). In relation to this, Rieh *et al.* identified a number of beneficiaries that were mentioned in the literature: authors are provided a service which includes long-term preservation and accessibility for their publications, readers get access to various types of material, and the institutions benefit from increased visibility for the work produced by their staff and students (Rieh *et al.*, 2008, p. 170). A number of other motivations were also mentioned in their interview study, and it was noted that some institutions indeed lacked clearly defined purposes for maintaining a repository (Rieh *et al.*, 2008).

Whereas open access is generally a clearly stated purpose with the repositories, several authors claim that preservation is explicitly mentioned less often, and less regulated through repository policies (Termens, *et al.*, 2015; Rieh *et al.*, 2008; Research Information Network. *Working Group...*, 2012;). Chowdhury (2014, p. 122) sets the percentage of repositories with a preservation policy to 8.1% based on figures in OpenDOAR.

Open file formats and open standards

Over the years, a number of file formats have been developed and implemented in software. For long-term preservation purposes it is critical that files created in a specific file format can be interpreted independently of the software used to create the file, since files typically outlive the software used to create it (Lundell, 2012). Previous research has shown that maintenance of software beyond a decade is a major challenge (Lundell *et al.* , 2011), and given that files often need to be preserved over a considerably longer period of time, it is essential that the file formats used are open file formats (Lundell, 2012).

A defining property of open file formats is that they can be implemented in software provided under different (proprietary and open source) licenses. A number of open file formats have been recommended for use in public sector organizations (e.g., Kammarkollegiet, 2016). Such file formats minimise the risks for undesired dependencies on specific proprietary technology controlled by specific vendors. For instance, research has shown that major challenges can be encountered when it comes to obtaining all necessary rights for implementing specific file formats in software (Lundel, Lings and Syberfeldt, 2015) and as a consequence such software cannot (legally) be used. Furthermore, even if all necessary rights can be obtained for implementation of a file format in software, research has shown that it constitutes a major challenge to maintain such software over the files' full life-cycle (Lundell, 2012). In such a scenario, an institutional repository would be unable to access and interpret its own content.

There is a complex relationship between the technical specification of a file format and its implementation in software, which imposes major challenges from the perspective of digital preservation. For example, for the PDF file-format it has been shown that some software implement features beyond the technical specification of the file format and that some software implements only a subset of the features in the technical specification of the file format (Gamalielsson and Lundell, 2013). This implies major challenges for the longevity of files since there is an inherent dependence on the software used to create each file over the full life-cycle for each file.

Some file formats have been recognised by standardisation organizations (and published as *standards*), whereas other formats are maintained by specific companies. For example, PDF/A-1 is a file format recognised by the International Organization of Standardization as an international standard, whereas TIFF 6.0 is a file format developed and maintained by a specific company (Adobe Systems Incorporated). Further, PDF/A-1 is recognised as an open standard according to the definition used in the Swedish public sector by Kammarkollegiet and included in their list of open standards recommended for use in public sector organizations (e.g., Kammarkollegiet, 2016). On the contrary, TIFF 6.0 is not recognised as an open standard under the same definition.

The interviews conducted in the study by Rieh and colleagues (2008, p. 178) indicated that some organizations purposefully choose open-source software for their institutional repository. A number of motivations were given for the choice, including an institution-wide ambition to implement open source software, that the software had been widely adopted, that it provided institutional control over the system, the low cost, and good functionality of the system (speed, flexibility and a good interface). In many cases, however, lack of access to technical support was a reason to acquire a proprietary system.

File stewardship in institutional repositories

A few studies with aims that partly overlap with ours provide a possibility for international comparison of results with a Spanish (Termens, *et al.*, 2015), an Indian (Sawant, 2011) and a US (Rimkus *et al.*, 2014) context. Termens and colleagues (2015) investigated the control of file formats in two large Spanish repositories from a preservation perspective. One was an institutional repository and the other an electronic dissertation repository. Unlike in the present study, Termens and colleagues harvested all files from the repositories and analysed them to determine file formats and encryption used. The study thus shows actual use of file formats rather than guidelines and attitudes.

In the doctoral thesis repository, the majority of files were PDF files (91.5 %) with most of the remaining files in JPEG (Termens, *et al.*, 2015, p. 165). In the other repository, which contained a broader variety of genres from one university, PDF was the main, primarily text-based, format (28 %), whereas image file formats were the most common file type (33 %) (p. 167). In both repositories, PDF versions 1.4 and 1.6 were the most common (accounting for 70% and 79%, respectively), and PDF/A was barely present at all. The PDF files were also often encrypted to prevent manipulation, which the authors see as a potential hindrance for future file migration. The authors conclude that there seems to be more focus on metadata than on file control in the repositories and that this can have consequences for long-term preservation of the repository objects.

Sawant (2011) used a survey to investigate various aspects of software implementation, supported file formats and preservation techniques in institutional repositories in India. DSpace was the most commonly used system (in eleven of fourteen institutional repositories, or 78%), which shows that DSpace was even more common than in a more global perspective, where Chowdhury found that 41.5% of repositories listed in OpenDOAR used DSpace (Chowdhury, 2014, p. 122). The repositories in Sawant's study all supported text file formats (e.g., HTML, PS, PDF, spreadsheets) and most supported image file formats. Slightly more than half of the repositories had support for audio and/or video formats, and a few supported datasets, databases, computer programs and CAD/CAM files. However, at the time of the study, the repositories only had text and image files deposited.

Rimkus *et al.* (2014) investigated file format policies and how much confidence American libraries (members of the Association of Research Libraries) placed in file formats, as stated in the policies. About half of the repositories and digital libraries identified had a file format policy, which is comparative to the occurrence of file format policies in the small set of policies studied by Probets and Jenkins (2006) in the UK, USA, Australia and Hong Kong, and lower than in the international study by Hitchcock, Brody, Hey and Carr (2007). In the study by Rimkus *et al.* (2014), a total of 174 file formats were mentioned in the policies, most of the type text or document. The most commonly mentioned file format was TIFF (in 115 policies), which was also the one most often considered highly trusted. It was followed by WAV (80), PDF (74), JPEG (70) and JPEG 2000 (68), Plain text (69) and Quicktime (67). Non-proprietary file formats were generally more trusted than proprietary ones. The authors found that format types and file formats that libraries had experience in curating were trusted more; this accounted for document/text and image files, primarily, whereas, for instance, computer programs, applications, video and tabular data were less highly trusted. Furthermore, many repositories relied on policies associated with particular repository or preservation software, rather than creating policies from scratch. However, the authors also identified that many repositories were willing to accept file formats used by their constituency of researchers regardless of how highly the repository trusted the formats, even though they only promised bit-level support of the files.

In a survey conducted by Hitchcock *et al.* (2007) involving twenty-one large repositories in Europe, the USA and Australia, PDF was commonly referred to as the preferred file format, but without details on precisely which version. Furthermore, repositories did not generally mention specific software applications, specific implementations of specific file formats and specific file formats controlled by specific suppliers. For example, unclear formulations from repositories included ‘PDF', which is imprecise, and ‘Adobe pdf', which is vague and may refer to a specific implementation of a specific file format or alternatively the documented technical specification of a specific version of a ‘pdf' file format. Fifteen repositories in the survey sometimes transformed files on or after submission, primarily to PDF. One repository mentioned conversion to PDF/A, however without specifying which version. Few respondents indicated that they required or encouraged authors to submit the source or initial file (e.g., in word processing format) for preservation. The authors identified risks involved in implementing policies that restrict the use of file formats, arguing that such conduct can lead to undocumented changes being made to the initial files, which in turn can involve loss of data. We note that the survey results reported in Hitchcock *et al.* (2007) show considerable confusion concerning differences between file formats as documented in technical specifications and software implementations of specific file formats, something which has later been observed in many public sector information technology projects conducted in and beyond the archiving domain, as reported in a study published by the Swedish Competition Authority (Lundell, Gamalielsson and Tengblad, 2016).

Method

This study investigates how management and staff at institutional repositories at Swedish higher education institutions express their views on the functions and life span of their institutional repository and describe how files and file formats are managed in the repositories. A questionnaire which allowed for both closed and open answers was deemed a suitable tool to gather data which could both provide a condensed picture of the repository landscape and offer participants the possibility to provide slightly more in-depth answers.

An online questionnaire consisting of thirty questions was developed and a link to it distributed to the thirty-five Swedish higher education institutions with an institutional repository. The institutions were those included on the list from the Swedish Higher Education Authority (2015) of the country's forty-eight higher education institutions. Thirteen of these do not have an institutional repository. The invitation to participate in the survey was distributed through an e-mail list for institutions that are part of the DiVA network. DiVA is a repository platform developed by Uppsala University Library (DiVA, n.d.) and used by twenty-eight of the institutions. Invitations to the remaining institutions were sent to e-mail addresses associated with each repository. The first round of invitations went out in early July 2015, just before summer vacation in Sweden. A reminder was sent in late August. A third reminder was sent to specific managers of repositories that had not yet replied in mid-September. Answers were received from all but one institution, with a resulting response rate of 97%.

Both closed and open questions were included in the questionnaire, and almost all questions allowed the respondent to comment on the question, an opportunity which was used in a few cases. The questions were categorised as background questions (type and size of the repository, types of content in the repository and views on the function of the repository), questions about file formats and file management, questions about versions and about rights management and questions about file depositing and the life-span of the repository.

The questionnaire was constructed using the tool Sunet Survey, which is provided by the Swedish university network. The data were exported to Excel and a report summarising the survey results has been produced using Sunet Survey's internal report manager. Because of the high response rate, the answers have been considered to provide good representation of the population of Swedish higher education institution repositories.

The closed question data were analysed in Excel using descriptive statistics. Open question data were coded to identify types of answers as well as number of similar answers. In a few instances, the answers to an open question showed that a closed question had been misinterpreted, in which cases the answers to the open question were given precedence in the analysis. In cases where this happened, it has been commented on in the presentation of the results.

One possible limitation to the study design is that the higher education institutions are represented by their institutional repository organizations (generally a library). These are primarily focused on open access publishing, which often means a focus on accessibility rather than on preservation. Preservation may be the task of other parts of the organization, such as the archive. The answers were provided by an individual working with the repository, although there are indications that several people collaborated on the answers in some cases. The invitation indicated that the best person to fill out the questionnaire was the *‘system owner/manager or other person working actively with the institutional repository'*. This means that the organization's policies and plans have been interpreted by one or a few people, but that the intention has been that this is a person well acquainted with those policies and plans. The higher education institutions included in the study were, furthermore, restricted to Sweden. These higher education institutions in most cases have a well-developed infrastructure for institutional repositories.

Results

Repository content

The questionnaire contained a number of questions about the content in the repositories and how it was managed, related to the first research question. The focus was on policy rather than on the actual deposits, and questions included which publication and format types as well as file formats were accepted, how file formats were treated and what life-span the respondents were expecting the repository and files to have.

Number of files

The number of research-related files (i.e. not student theses) uploaded during 2014 in the repositories that participated in the questionnaire ranged between none and 3,626, with a total of 6 repositories listing more than 1,000 files deposited. There have been some difficulties in a few cases for the respondents to produce data that distinguish between files uploaded in the repository during 2014 and files with the publication date 2014, as well as distinguishing between the number of open access publications and those files deposited that are not available open access. The corresponding figures for student theses range between 10 and 3,560, with 9 institutions listing 1,000 files or more and one institution answering ‘don't know'. Those who comment on the figures note that the changes compared to previous years are negligible, or that they can observe a slight rise in the figures over the years. The rise is attributed to increased awareness of open access among researchers or to mandates to deposit student theses in the repository.

Publication types

A number of publication types were accepted by the repositories (see Figure 1). All but one repository accepted journal articles and all but one repository accepted dissertations. Almost all repositories accepted book chapters, published conference papers and posters, reports and publication series, reviews and books. Fewer than half of the repositories at the time accepted digitised material, research data and open learning resources. No question was posed about student theses, but based on the answers to the question about the number of student theses deposited it can be assumed that all repositories also include student work.

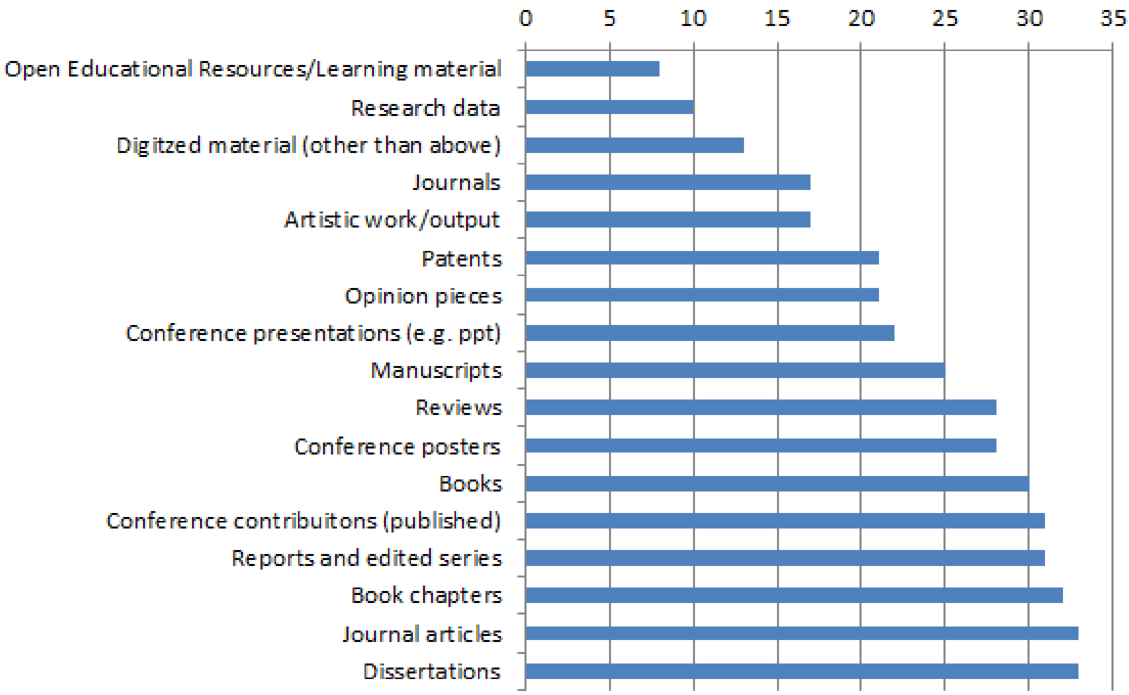

Figure 1: Publication types accepted in the repositories (n=34).

Format types

The respondents were asked about which format types they accepted in the repositories: text, image, audio, video and multimedia. Slightly more than half (19) accepted all these types, and an additional five, all but multimedia (although there was some uncertainty about how multimedia should be interpreted in this case). Seven repositories only accepted text files. This dominance of text but with a fairly broad acceptance of other format types mirrors the findings in Sawant (2011) from Indian repositories and of the Spanish doctoral thesis repository (Termens, *et el.*, 2015).

Accepted file formats

About 70% (24) of the repositories stated that they had some form of instructions or policy concerning which file formats were accepted in the repository. This is slightly higher than among the American repositories investigated by Rimkus *et al.* (2014) and those surveyed by Hitchcock *et al.* (2007), although the difference could depend on the question in our survey being phrased to include a bit more than formal policies, or on the time that has passed since the 2007 survey. A few of the repositories that lacked instructions mentioned that there was no need for restrictions ('*"everything" is accepted*'). It was also mentioned that departments may have special instructions for student theses. Eight of the repositories referred to the extensive list of accepted file formats in DiVA, although half of them had interpreted the list as instructions and half had not. DiVA, at the time of writing, accepted the following file formats: text (xml, txt, pdf, ps, csv, epub); image (tiff, png, jp2, jpg); audio (wav, mp3); video (mpeg, mov, flv, avi, msvideo, mp4 [mp4, 3gpp]); compressed formats (zip, gz, kmz); and geographic data (kml) (according to the form). It should be noted that for several of these formats (e.g., pdf, tiff, jp2 and mov), DiVA does not specify which specific version (or versions) of a specific file format are accepted. Such reliance on platform-associated policies or lists was also similar to findings by Rimkus *et al.* (2014) and Hitchcock *et al.* (2007).

Among those who mentioned specific file formats, most (12) had recommendations to use PDF files, although they also accepted other formats or mentioned that the recommendation primarily concerned text files or student theses. One respondent mentioned that mp3 was recommended for audio and mp4 for video. However, only on one occasion was a particular version of the file format named. Three repositories gave answers that in some way connected to the issue of open file formats and/or long-term preservation. Blekinge Institute of Technology accepted file formats that can be opened through free software. Linnaeus University mentioned that they accept sustainable or constant formats, and the University of Skövde, where two of the authors work, mentioned a requirement for '*at least PDF/A-1b'*, at least for doctoral dissertations.

Checking and converting files

Around 70% of the repositories also stated that they check that the deposited files adhere to the guidelines. The comments indicated that these checks had less to do with file formats than with other aspects of the file, but some mentioned that there was a system check to ensure that the file format was the correct one, and that the library staff opened the file to make sure that it worked. Others mentioned that the file was checked for copyright issues (for example through SHERPA/RoMEO) and in one case that staff attempted to secure that the version of the publication was the one reported.

Some of the repositories stated that they sometimes (though not often) converted files from one format to another, which was also the case in the study by Hitchcock *et al.* (2007). In almost all cases when this was described in our study, it was a matter of converting from DOC to PDF, either as a service to the author before depositing the file or when files had been uploaded, as part of system migration, or when older files were encountered in the database that did not follow policy, since DiVA previously accepted the uploading of DOC files. A couple of repositories mentioned that large files may be compressed to ZIP. Many of the repositories had not experienced any need to convert files.

Open standards and file formats

Discussions about the use of open standards and open file formats in the repositories were not running high at the institutions. Three quarters of the respondents stated that there were no such discussions, or that they were not aware of any. In the comments, some referred to discussions that had been and were taking place in the DiVA consortium and that aimed to only accept stable file formats to be uploaded in the system in order to ensure long-term preservation. At the same time, another respondent highlighted that there are advantages to accepting all file formats used by the authors. In personal communication, the project leader of DiVA confirmed that although there is no strict policy, the ambition has been to support stable formats that will be sustainable over time. However, he mentioned that, when the platform begins to be used for research data sets in the near future, a wider set of formats will be allowed (personal e-mail from Urban Ericson, 2016-01-08). One respondent mentioned that the university archive (another division of the institution) was investigating open file formats, and one repository stated that they discuss the use of PDF/A.

Depositing ‘hidden' files in the repository

Almost all respondents (31) stated that files were sometimes deposited in the repository without being made openly available immediately. All repositories used systems that allowed for information about embargo periods to be registered. The few comments on this question indicated that keeping files deposited but not openly available in the repository happened for two reasons. Firstly, with student theses and doctoral dissertations where depositing in the repository was regulated through policy, but where the file, for one reason or another, could not be made openly available immediately or would be made available on a specific date. Secondly, as a temporary action in the case of parallel publishing of publications that were to become open access after an embargo period (this is required, for instance, in the case of Horizon 2020-funded projects). Although some publishing policies for the higher education institutions stated that student theses and doctoral dissertations should be deposited in – and if possible also made open access through – the repository, there was no mention of attempts to use the repository as a tool for institutional access to not only metadata about, but also to the full-text of, publications authored by employees. However, one respondent mentioned that a purpose of the repository was to provide private access to files for authors.

Repository life-span

When asked about how long-term the plans for the repositories and the files in them are, the answers primarily indicated uncertainty. Around 60% of the respondents said they do not know how long-term the plan is for the repository or for the life of the stored files. However, those that do have a plan looked towards the future, and indicated more than twenty years for the life of the repository (7) and for the life of the files (13, with an additional 2 commenting that they work with an ambition of more than twenty years, but do not have an agreed upon plan). Five respondents answered that the repository had a fairly short-term plan of less than five years. In one case, however, this referred to an operational plan which was updated each year, thus indicating that the plans for the life of the repository may be longer. In another case the answer referred to a planned systems change. When commenting on the question, respondents noted that it is difficult to predict the future, and that we do not know what academic publishing will be twenty years from now. They also brought forth that the accessibility of files is likely to differ between file formats, but that there will probably be ways to migrate common file formats such as PDF, if needed. One respondent would prefer to see files stored as HTML in order to ensure long-term preservation.

Views on repository functions

One aspect which potentially influences if and how repositories plan for the future and which file formats they choose to accept is which current and future functions they consider to be important. In order to understand how the repository organizations view repository functions, respondents were asked to rate the importance of various functions on a five-grade scale (see Table 1). The questions addressed some of the functions mentioned in the literature, such as access to and preservation of documents.

The answers to these ratings showed that visibility of research publications and student theses was what most respondents agreed was an important function for the repository (see Table 1). This was followed by the role of the repository as a primary publication site for research that is generally not published elsewhere, such as dissertations and student theses, and for securing access to the institution's research publications. A few smaller higher education institutions did not agree that primary digital publication is the role of the repository, but the majority of participants considered it an important function. Almost all respondents also partly or fully agreed that a function for the repository was to provide a site for depositing files in order to facilitate for employees to adhere to open access mandates. Few Swedish higher education institutions have strong institutional mandates, but many have policies that encourage their employees to publish open access (Francke, 2013). However, the fact that many of the major research funders have mandates in place seems to at least partly motivate the high agreement on this question, and perhaps makes the many partial agreements a bit surprising.

While there was a fairly agreed view that visibility and access to research publications and student theses were activities the repositories should work with, the answers were more scattered when it came to long-term preservation of publications. This finding corresponds to results from a survey involving UK institutional repository managers (Jones *et al.*, 2008, p. 13). The publication type that most respondents (a little more than half) agreed should be preserved long-term in the repositories was student theses. An explanation may be found in the fact that some of the Swedish higher education institutions no longer store paper versions of student theses but, rather, deposit digital copies in the repositories. In some cases, this copy is considered the archived copy (see ‘Depositing hidden files in the repository' above and Francke, 2013), which motivates long-term preservation. In response to another question, fourteen respondents mentioned that storage of files, such as student theses or doctoral dissertations, was a reason for depositing them. Storage was either intended to be long-term or temporary for future transfer to the institution's archive (7) or while the institution decided how to store them long-term elsewhere (10). Regarding the repository's functions, more respondents fully agreed that student theses were important to preserve long-term than was the case with research publications. In the latter case, respondents were a bit more likely to indicate that they partly agreed. In most cases, the institutions that fully disagreed took the same stance in relation to long-term preservation of all publication types, with one exception for student theses. On another question, two respondents specifically highlighted that one of the purposes with the repository was long-term preservation of files.

Question	1	2	3	4	5	Avg.
Visibility of institution's research	-	-	-	3	31	4.9
Secure access to institution's research	-	-	7	8	19	4.4
Long-term preservation of institution's research publications	5	-	5	10	14	3.8
Primary publication site for e.g., dissertations	3	-	2	3	26	4.4
Visibility of institution's student theses	-	-	1	3	30	4.9
Secure access to institution's student theses	-	1	1	7	25	4.6
Long-term preservation of institution's student theses	4	-	4	7	19	4.1
Secure access to institution's research data	13	4	8	7	2	2.4
Long-term preservation of institution's research data	12	5	8	6	3	2.5
Facilitate employees' adherence to open access mandates from e.g., funders	1	-	4	11	18	4.3
Generate traffic to institution's Website	2	3	11	9	9	3.6

Table 1: The extent to which the respondent agreed with the claim that these are important functions for the institutional repository (n=34). 1 (fully disagree) to 5 (fully agree)

During the time leading up to and following the data collection, there has been much discussion at Swedish higher education institutions and their libraries about open access to, and preservation of, research data. The discussion was intensified by the work leading to the *Proposal for national guidelines for open access to scientific information*, which was submitted to the government in January 2015 and which included guidelines for both publications and research data (Swedish Research Council, 2015). Most institutions seem to foresee that Sweden will follow international initiatives for open access to research data, but there is, at the time of writing, not yet a strategy or system in place for access to and preservation of data. This motivated asking about the view of research data in the repositories. The answers show that it is quite clear that research data do not hold the same established position in relation to the repositories as do publications, even if some repositories stated that they accepted research data as deposits (see Figure 1). The majority of the respondents fully or partially disagreed with the statement that a function for the repository should be to secure access to or provide a site for long-term preservation for research data. Unlike for publications, the views on access and preservation are very similar. It could also be noted that about a quarter of the respondents do view access and preservation of research data as a function for the repository and that about as many are neutral. Even though it was the least agreed upon function, it was not universally dismissed.

Closely related to visibility, but at least potentially associated with marketing values not originally expressed by the open access movement, is the wish to direct traffic to the institution's Website. This was the question where most respondents, about one third, chose the neutral option. At the same time, most of the respondents, about two thirds, at least partly agreed that this was a function for the repository.

In 2015, a complement to the Legal Deposit Act (1993:1392) that concerned deposit of electronic materials came into effect in Sweden (National Library of Sweden, 2014). A few respondents mentioned in comments that their institutions were planning to use the repository for file delivery in order to fulfil their legal obligations.

Furthermore, a number of respondents (9) agreed that the repository was also used for statistics and various types of analyses. A few specifically mentioned that the repository served as a search portal for the institution's publications (although this could also be attributed to the metadata only).

We were curious to see if we could detect any patterns among the institutions based on their answers to the questions about the repository's function in Table 1. However, neither manual analysis nor hierarchical clustering showed patterns among the responding institutions or repositories that we could identify as meaningful. On the contrary, the clusters that emerged were highly diverse with regard to such factors as size and specialty of the institution, geographical location, repository platform, and open access mandates.

Discussion

At the beginning of this paper we posed three research questions. The first two, "How do Swedish institutional repositories restrict, promote, manage and document file formats of files deposited in the repositories?" and "How do respondents at Swedish institutional repositories describe the function of the repositories?" have been addressed above. In the following, we consider the third question which brings the two previous ones together: "What are some expected consequences of the answers to these two questions for the content stewardship in the institutional repositories to ensure the sustainability of the content?" The question connects to the task set out for institutional repositories by Clifford Lynch to provide '*an organizational commitment to the stewardship of... digital materials, including long-term preservation where appropriate, as well as organization and access or distribution'* (2003, p. 2).

We approach this question from the assumption that the repositories can be viewed as a type of infrastructure, and that infrastructures occur as part of practices (Star and Ruhleder, 1996). Consequently, *when* an infrastructure takes place is important; it does so in relation to certain activities, organizational settings, politics, norms and technologies. This perspective has been developed by Star and Ruhleder (1996) and further interpreted by Bowker, Baker, Millerand and Ribes (2010) as a distribution in two dimensions based on the axes local-global and technical-social.

The answers to the question of which functions the repositories are considered to have are closely associated with the repository's position in the organization, but also with the various communities that provide the files deposited in the repository. Views of the repository's functions, and choices made with regard to the repository, are thus embedded in local library, management, research and teaching organizations, norms and expectations. They are shaped by and co-shape views of publishing and communication, including openness in communication, in disciplinary and library practices.

It can be argued that a collecting function influences the view of the repository and, thus, also the choices made with regard to file formats. The wish to get material into the repositories has often been more important than ensuring that the files are in formats that are suitable for preservation. The ambition to collect material is understandable, as a repository without material will remain unused. It was mentioned that measures were sometimes taken to accommodate the preferences of users who upload files. Similar findings, showing that repositories would accept the formats offered by researchers, were found for US libraries by Rimkus *et al.* (2014). Thus, the focus on collecting material is closely associated with, or even disguised as, a user perspective. Many repositories also mention that they open the files to make sure that they are functional, which indicates a concern with accessibility if not with preservation. Visibility and access as core functions for the repositories were emphasised in the responses to the questionnaire.

However, a move towards restricting which file formats are accepted has taken place within the DiVA consortium. A large number of repositories (about 70%) have an instruction or policy specifying which file formats are accepted, although the policies rarely seem to advise on particular versions of the file format (see below). The restriction of file formats to exclude certain formats that are not considered to be stable can be interpreted in terms of adjustment not to the (local) providers of files and, by extension, to their respective communities, but rather to the broader discussion of openness within the open science movement which is very active within much of the repository community. At the same time, there were very few indications in the replies that discussions about open standards and file formats take place at the institutions (75% were not aware of any such discussions). Furthermore, among the many file formats accepted in DiVA, some but not all meet the requirements for being open file formats. One of the challenges faced by the repositories, which is not likely to diminish in the future, is the tension between, on the one hand, the use of a wide variety of closed file formats embedded in the social arrangements, and technology use of local file producer communities (employees and students), and on the other hand the embodiment of standards through file formats and software that are predictable and interoperable and which thus entail a higher possibility for replication in the future. This tension is distributed along the axes of social, institutional enactment of file production and its technical configuration in more or less standardised and transparent protocols (see Bowker *et al.*, 2010, p. 101).

The publication types and format types accepted by the repositories were often quite varied, and even though types accepted by all or most repositories are primarily text-based, the variety involves additional demands on the stewardship of file formats, including for the seventeen repositories that accept artistic works or output. Some of these publications can also be expected to be made 'gold' open access through the repository, for instance dissertations and student theses. The repository's function as a primary publication site for these publications was also viewed as at least fairly important by all but a few respondents. Responses indicated that there were sometimes specified requirements when it came to file formats for these publications, such as institutional policies for student theses. More than 70% of respondents also agreed or partly agreed that long-term preservation of student theses and research publications were important tasks for the repository. In some cases, the repository was viewed as a temporary storage space while investigations into more long-term solutions took place and a few repositories listed long-term preservation as an explicitly stated function. In fact, about one fifth of repositories planned for a life-span for the repository exceeding twenty years, and an additional fifth for the same life-span for the files if not the repository.

Thus, the answers revealed a consideration of the content stewardship in the repositories. In some cases, this also took the form of converting files into what was considered more sustainable file formats, as from DOC to PDF. However, there was little indication in the responses that attention was paid to which version of the file formats (in particular PDF) was used (see also Hitchcock *et al.*, 2007). File format versions were not specifically asked about in the questionnaire, because we wanted to avoid steering the answers too much. In hindsight, this may have been a mistake. However, when asked to reflect on conversion of files, which formats files were converted into and on discussions about open file formats, only very few of the respondents mentioned the versions of formats used. This may be an indication that little distinction is made as part of the repository work between various versions of, for example, PDF. This may in turn be problematic in light of the long-term preservation and accessibility of the repository content, given that Lundell *et al.* (2011) have shown that the longevity of software which is not open-source is a major challenge. Furthermore, because the longevity of software does not meet requirements for life-cycles for preservation of files, it is important that file formats in different versions are open file formats if preservation is to be ensured beyond a decade.

Conclusions

The institutional repositories at Swedish higher education institutions do not, in most cases, have an archival responsibility for any of the publications deposited in them. According to the answers provided by the repository managers and staff in this study, visibility and secure access to the content in the repositories are their primary functions. At the same time, most of the respondents at least partly agree with the statement that long-term preservation of some publication types is an important function for the repository and that access is not only a current but also a future concern. In light of this, there may be reasons for the repositories and their host institutions to more carefully consider their policies and guidelines on file formats in order to support the sustainability of the repository and the future accessibility of its content. This can be achieved by advising or requesting authors to deposit files not only in a particular file format, but in a version of the file format which adheres to an open standard. Provided that the main responsibility is placed on the depositor rather than on repository staff, this would require an attempt to change scholarly and organizational practices through guidelines, instructions and advice. This need will arise especially as we move beyond the most common text-based file formats.

An alternative strategy, which is potentially more work-intensive for the repositories, would be to migrate files and perform conformance checking of files in different file formats as part of the process of checking metadata and files when they are deposited, work which is already carried out to some extent. Changing file format practices of various disciplinary communities is not necessarily an easy task and will have different implications in different communities. The task is further complicated by the unpredictable ways in which much software produce files in specific formats, even in versions that can be expected to follow open standards. Despite the challenges involved, instructing authors to deposit files saved as PDF/A-1 or PNG, where possible, can be a first step towards changing practices which can, with time, lead to repositories that will be more likely to be accessible and open not only today but in five or twenty years. It is time that the issue of open and standardised file formats – of sustainable open file formats – becomes a focus of attention for institutional repositories in order to avoid problems of accessibility in the near future.

Acknowledgements

We would like to give warm thanks to all Swedish institutional repository managers and staff who took the time to provide extensive answers to the questionnaire. A special thank you to Pieta Eklund and Thomas Nyström at the University of Borås Library for comments on the first draft of the questionnaire and to Urban Ericsson and his team at the Uppsala University Library for help with distributing the invitation to participants and for extracting data to help the participating DiVA libraries to answer some of the questions. The study was funded by Region Västra Götaland, the University of Borås and the University of Skövde.

About the authors

Helena Francke is an Associate Professor of Library and Information Science at the Swedish School of Library and Information Science, University of Borås, Sweden. She can be contacted at Helena.Francke@hb.se
Jonas Gamalielsson is a researcher at the School of Informatics, University of Skövde, Skövde, Sweden. He can be contacted

atjonas.gamalielsson@his.se

Björn Lundell is a Associate Professor at the School of Informatics, University of Skövde, Skövde, Sweden. He can be contacted at bjorn.lundell@his.se

References

Ball, A. (2010). *Preservation and curation in institutional repositories* (version 1.3). Edinburgh, UK: Digital Curation Centre. Retrieved from http://1seminariopreservacaopatrimoniodigital.dglab.gov.pt/wp-content/uploads/sites/19/2015/08/recurso_02.pdf [Unable to archive]

Björk, B.-C., Lakso, M., Weiling, P. & Paetau, P. (2014). Anatomy of green open access. *Journal of the American Society for Information Science and Technology, 65*(2), 237-250.

Bowker, G.C., Baker, K., Millerand, F. & Ribes, D. (2010). Toward information infrastructure studies: ways of knowing in a networked environment. In J. Hunsinger, L. Klastrup & M. Allen (Eds.), *International handbook of Internet research* (pp. 97-117). Dordrecht, the Netherlands: Springer.

Cho, J. (2014). Intellectual structure of the institutional repository field: a co-word analysis. *Journal of Information Science, 40*(3), 386-397.

Chowdhury, G.G. (2014). *Sustainability of scholarly information*. London: Facet.

DiVA – Digitala Vetenskapliga Arkivet (n.d.). About DiVA portal. Retrieved from http://www.diva-portal.org/smash/aboutdiva.jsf?dswid=ppfmain (Archived by WebCite® at http://www.webcitation.org/6qWy9Xqxs)

Eschenfelder, K.R. & Shankar, K. (2016). Designing sustainable data archives: comparing sustainability frameworks. In *iConference 2016 Preliminary Results Papers* (7 pages). Urbana, IL: University of Illinois, IDEALS. Retrieved from http://bit.ly/2qtsffj (Archived by WebCite® at http://www.webcitation.org/6qX4v2Qat)

Francke, H. (2013). *Publicera! Svenska forskningsbiblioteks arbete med publiceringsfrågor* [Publish! Swedish research libraries' work regarding publishing]. Stockholm: Swedish Library Association.

Gamalielsson, J. & Lundell, B. (2013). Experiences from implementing PDF in open source: challenges and opportunities for standardisation processe. In K. Jakobs (Ed.), *Proceedings of the 8th IEEE Conference on Standardization and Innovation in Information Technology (SIIT 2013)*(pp. 39-49). Piscataway, NJ: IEEE.

Hitchcock, S., Brody, T., Hey, J.M.N. & Carr, L. (2007). Survey of repository preservation policy and activity . Southampton, UK: University of Southampton. Retrieved from http://preserv.eprints.org/papers/survey/survey-results.html (Archived by WebCite® at http://www.webcitation.org/6m3bkV00T)

Jacobs, N., Thomas, A. & McGregor, A. (2008). Institutional repositories in the UK: the JiSC approach. *Library Trends, 57*(2), 124-141.

Jones, C., Darby, R., Gilbert, L. & Lambert, S. (2008). *Report of the subject and institutional repositories interactions study*. London: JISC & Science and Technology Facilities Council. Retrieved from http://repository.jisc.ac.uk/259/1/siris-report-nov-2008.pdf (Archived by WebCite® at http://www.webcitation.org/6ocdQCRy2)

Kammarkollegiet. (2016). *Öppna standarder: programvaror och tjänster 2014* [Open standards: software and services 2014]. Stockholm: Kammarkollegiet, Statens inköpscentral. (Dnr 96-38-2014). Retrieved from http://www.avropa.se/contentassets/8d843fb85c8f40ab9ba5c4acc2d1ecfc/oppna-standarder---programvaror-och-tjanster.pdf (Archived by WebCite® at http://www.webcitation.org/6qX5V2q3A)

Kennan, M.A. & Wilson, C. (2006). Institutional repositories: review and an information systems perspective. *Library Management, 27*(4/5), 236-248.

Kunda, S. & Anderson-Wilk, M. (2011). Community stories and institutional stewardship: digital curation's dual roles of story creation and resource preservation. *portal: Libraries and the Academy, 11*(4), 895-914.

Lundell, B. (2012). Why do we need open standards? In Marta Orviska & Kai Jakobs (Eds.), *Proceedings of the 17th EURAS Annual Standardisation Conference 'Standards and Innovation'* (pp. 227-240). Aachen, Germany: G. Mainz Verlag

Lundell, B., Gamalielsson, J. & Katz, A. (2015). On implementation of open standards in software: to what extent can ISO standards be implemented in open source software? *International Journal of Standardization Research, 13*(1), 47-73.

Lundell, B., Gamalielsson, J. & Tengblad, S. (2016). *IT-standarder, inlåsning och konkurrens: En analys av policy och praktik inom svensk förvaltning [IT standards, lock-in and competition: an analysis of policy and practice within Swedish administration]*. Stockholm: Konkurrensverket. (Uppdragsforskningsrapport, 2016:2).

Lundell, B., Lings, B. & Syberfeldt, A. (2011). Practitioner perceptions of open source software in the embedded systems area. *The Journal of Systems and Software, 84*(9), 1540-1549.

Lynch, C.A. (2003). Institutional repositories: essential infrastructure for scholarship in the digital age. *portal: Libraries and the Acacemy, 3*(2), 327-336.

National Library of Sweden (2014). *Legal deposits of electronic materials in Sweden*. Stockholm: National Library of Sweden. Retrieved from http://www.kb.se/dokument/Pliktleverans/Eplikt_myndigheter_eng140917.pdf (Archived by WebCite® at http://www.webcitation.org/6qX64KPuK)

OpenDOAR. (2014). *The directory of open access repositories*. University of Nottingham. Retrieved from http://www.opendoar.org/ (Archived by WebCite® at http://www.webcitation.org/6qX6E7PQ4)

Probets, S. & Jenkins, C. (2006). Documentation for institutional repositories. *Learned Publishing*, *19*(1), 57-71.

Research Information Network. *Working Group on Expanding Access to Published Research Findings*. (2012). Accessibility, sustainability, excellence: how to expand access to research publications. Research Information Network. Retrieved from https://www.acu.ac.uk/research-information-network/finch-report-final (Archived by WebCite® at http://www.webcitation.org/6hgyglgI2)

Rieger, O.Y. (2012). Sustainability: scholarly repository as an enterprise. *Bulletin of the American Association for Information Science and Technology*, *39*(1). Retrieved from http://asis.org/Bulletin/Oct-12/OctNov12_Rieger.pdf (Archived by WebCite® at http://www.webcitation.org/6qX6WEpkW)

Rieh, S.Y., St Jean, B., Yakel, E., Markey, K. & Kim, J. (2008). Perceptions and experiences of staff in the planning and implementation of institutional repositories. *Library Trends*, *57*(2), 168-190.

Rimkus, K., Padilla, T., Popp, T. & Martin, G. (2014). Digital preservation file format policies of ARL member libraries: an analysis. *D-Lib Magazine*, *20*(3/4). Retrieved from http://www.dlib.org/dlib/march14/rimkus/03rimkus.html [Unable to archive]

Robertson, W.C. & Borchert, C.A. (2014). Preserving content from your institutional repository. *Serials Librarian*, *66*(1-4), 278-288.

Sawant, S. (2011). IR system and features: study of Indian scenario. *Library Hi Tech*, *29*(1), 161-172.

Star, S.L. & Ruhleder, K. (1996). Steps toward an ecology of infrastructure: design and access for large information spaces. *Information Systems Research*, *7*(1), 111-134.

Stevenson, J.A. & Zhang, J. (2015). A temporal analysis of institutional repository research. *Scientometrics*, *105*(3), 1491-1525.

Swedish Higher Education Authority. (2015). *Higher education institutions (HEIs)*. Stockholm: Swedish Higher Education Authority. Retrieved from http://english.uka.se/higher-education-system/higher-education-institutions.html

(Archived by WebCite® at http://www.webcitation.org/6qX7Fz4rC)

Swedish Research Council. (2015). *Proposal for national guidelines for open access to scientific information*. Stockholm: Swedish Research Council. Retrieved from https://publikationer.vr.se/en/product/proposal-for-national-guidelines-for-open-access-to-scientific-information/ [Unable to archive]

Termens, M., Ribera, M. & Locher, A. (2015). An analysis of file format control in institutional repositories. *Library Hi Tech, 33*(2), 162-174.

How to cite this paper

Francke, H., Gamalielsson, J. & Lundell, B. (2017). Institutional repositories as infrastructures for long-term preservation. *Information Research, 22*(2), paper 757. Retrieved from http://InformationR.net/ir/22-2/paper757.html (Archived by WebCite® at http://www.webcitation.org/6r2RUhEeO)

Find other papers on this subject

Scholar Search Google Search Bing

Check for citations, using Google Scholar

Facebook Twitter LinkedIn Delicious More

Contents | Author index | Subject index | Search | Home