

Student Reidun Anette Hagen

“Digital Repositories”

An investigation of best practices for content recruitment to academic digital repositories and the conditions for their livelihood



Trondheim, the 25th of June, 2009.

Norwegian University of Science and Technology
Faculty of Social Sciences and Technology Management
Department of Industrial Economics and Technology Management



1 Preface

The paper you currently hold is my Master thesis on the topic of digital repositories. It is written as the contribution from the final semester of my master degree within Industrial Economics and Technology Management at the Norwegian University of Science and Technology (NTNU). The research documented by this thesis was conducted during my internship at the European Center for Nuclear Research (CERN), where I was part of the CERN Library team. My work at CERN has covered the measurement of quality in scientific journals and also an investigation of best practices for digital repositories in general, and for the CERN-Theses Database in particular. The latter has served as the inspiration and foundation for this thesis. This completed thesis and related work has been done in the time period stretching from the 15th of February until the 25th of June 2009, and counts for 30 ECTS points (studiepoeng).

At this point, I would like to thank my supervisor at NTNU, Roger Sørheim, who has kept me on track when I could not see the wood for the trees. I would also like to express my gratitude to the head of the CERN library, Jens Vigen, for giving me the possibility to experience CERN and for his invaluable advice during the development of this thesis.

Happy reading,

Reidun Anette Hagen

NTNU and CERN
June, 2009.

2 Abstract

A digital repository is a web accessible database, aimed at preserving the research material of an institution or scientific community. A digital repository serves as a tool for dissemination of research material and can increase the impact of the research by making it freely accessible. Digital repositories are often mentioned as a possible aid in relation to the Open Access debate; how research material should be freely accessible to anyone, anywhere at any time. However, for a digital repository to fully unleash its potential as a crucial component of Open Access, it is reliant on the ability to successfully collect and organize content. To a large extent this involves initiating self-archiving of research material by scientists throughout the academic world. This is not a trivial task, and many current repositories are inadequate in this respect, remaining empty, unvisited shelves. This thesis explores best practices for content recruitment to digital repositories, through the review of literature, and an empirical investigation of practices and experience from Electronic Theses and Dissertations repositories. The motives for self-archiving amongst users of the CERN-Theses Database are also assessed.

The thesis concludes that mandatory policies for self-archiving at institutions is to be recommended. Furthermore, this thesis suggests that the practice of self-archiving in institutional repositories should be initiated, rather than relying on alternative ways of recruiting content (for example archiving by library staff and administrative personnel). The work recommends merging the institutional repository with the institution's working tools for storing of work in progress and work sharing. Repository administrators must acknowledge the fact that many small repositories internal to various groups or sections may exist throughout an organization, representing a barrier to the adoption of a new, institution wide repository. Additionally, the thesis discusses the purpose of an institutional repository contra a subject repository as well as the appropriate level of co-existence between institutional repositories, subject repositories and journals, both subscription based and Open Access. The thesis argues that each of these entities serve a distinct function within the value chain of academic communication, and are thus suitable for co-existence. However, all stakeholders might benefit from clearer role division, and specialization towards their core functions. Institutional repositories are identified as best suited for recruiting content through mandatory policies for self-archiving, whilst subject repositories (by aggregating subject-related information from institutional repositories and offering extended search capabilities) are the bodies best suited for serving a scientist's need to find relevant information, as well as the dissemination of research results. Journals provide the much needed function of classification in terms of peer review, and may also offer prestige in terms of academic reputation to a scientist, and based on this analysis might do wise in focusing their business model towards this offering, rather than the dissemination and preservation of knowledge which might be better tended to by other bodies. Finally, the thesis discusses the future role of digital repositories in the context of Open Access, suggesting that we are more likely to achieve Open Access in the short run through mandatory policies for self-archiving in institutional repositories, rather than waiting for journals to find and adapt to viable Open Access business models.

3 Table of Content

1	Preface	i
2	Abstract.....	ii
3	Table of Content	iii
4	List of Figures	iv
5	List of Tables	iv
6	Introduction	1
7	Scope of Thesis and Lay Out	2
Section 1 – Theory Review		3
8	Digital Repositories	3
8.1	Definition Digital Repository	3
8.2	Their Purpose.....	4
8.3	The Extent of Digital Repositories	4
8.4	Interoperability across Repositories.....	5
8.5	Business Models for Digital Repositories	6
9	Repositories in a Context.....	8
9.1	Libraries and Repositories	8
9.2	Digital Repositories, Publishers and Journals.....	9
9.3	Open Access and Repositories.....	11
10	Content Recruitment	15
10.1	Strategies for Content Recruitment	15
10.2	Reasons for Poor Faulty Uptake	16
10.3	Success Factors	18
10.4	ArXiv – An Example of a Successful Digital Repository.....	18
11	Summary of Theory Review and Further Work	21
Section 2 – Method, Empirical Investigation and Discussion.....		22
12	Methodology	22
12.1	The Background for the Thesis	22
12.2	Literature	22
12.3	Empirics	23
13	Empirical Investigation	27
13.1	Research Questions	27
13.2	Findings from Investigation of ETD Repositories.....	27
13.3	Findings from Investigation of the CERN-Theses Database	29
14	Discussion	33
14.1	The Purpose of Different Digital Repositories	33
14.2	Best Practices for Content Recruitment.....	35
14.3	Co-Existence between Digital Repositories and Publishers	39
14.4	Digital Repositories and Open Access	42
14.5	Recommendation for the CERN-Theses Database	44
Section 3 - Conclusion and Further Research.....		46
15	Conclusion.....	46
16	Bibliography	49
17	Appendix	55
17.1	Glossary	55
17.2	The ETD Repositories in the Investigation.....	56
17.3	The Questionnaire sent by e-mail to the ETD Repositories	56
17.4	Templates for the interviews with users of the CERN-Theses Database	58
17.5	Original Scope of Thesis (Oppgaveteksten).....	61
17.6	Self-Declaration Form (Egenerklærings skjema)	62

4 List of Figures

Figure 1 - Repository Interoperability (From Jones, Andrew and MacColl 2006)	6
Figure 2 - Digital Repositories and Their Stakeholders	8

5 List of Tables

Table 1 – Summary of Reasons for Poor Faculty Uptake	18
Table 2 – The Purpose of Institutional Repositories Contra Subject Repositories.....	34
Table 3 – ETD Repositories in the Investigation	56

6 Introduction

It seems only natural that an academic institution or scientific community would want to capture and preserve its research output, as opposed to retrieving what can be assumed a fraction of its actual production by subscribing for journals. Journals with restricted access can also be seen as a barrier to efficient and far reaching dissemination of research which is a goal for a scientist who wishes to impact the future course of research within his discipline. Digital repositories are web accessible databases which aid preservation and dissemination of an institution's or community's research output. However, for this to be successful, repository administrators are reliant on author participation in terms of self-archiving in order to capture the large amounts of research output. This factor has proven itself as far from trivial, and faculty participation remains low in most disciplines (Lynch and Lippincott 2005, van Westrienen and Lynch 2005). Almost every academic institution in Europe and North America has got their own institutional repository by now (Swan 2008) and large amounts of resources have been put towards the creation of the digital repositories, as well as the continuous expenditures for running them. Despite all the resources used, many remain nearly empty and/or unvisited.

As the majority of research is funded through public money (Guédon 2004), it can be argued that the public should be able to access these results at a later stage, without paying a commercial third party in order to do so. Nor should research and science become a privilege reserved only for rich countries in the western world. Due to this reasoning, and the pressure resulting from inflating subscription prices over the previous decade (Swan 2006), the Open Access movement which works for free access to scientific literature, to anyone, anywhere at any time, has gathered momentum. Digital repositories are frequently mentioned in the context of Open Access, as a possible option for achieving it. However, this is not a conflict free or well painted out path. Some argue that digital repositories will be the end of journals that are the bodies providing the much wanted quality assurance (peer review) of research material. Whilst others again argue that a symbioses of journals and repositories are possible.

For digital repositories to fully unleash their potential and not be a mere resource drain with poor returns, the challenge of recruiting content to them must be addressed and investigated, which is what this thesis will do. Furthermore the thesis will discuss the livelihood of repositories in co-existence with the other actors within the value chain of scholarly communication.

7 Scope of Thesis and Lay Out

By review of repository related literature and an empirical investigation this thesis seeks to develop new insight to the following matters:

- **The purpose of different digital repositories**
 - o What is the purpose of an institutional repository contra a subject repository?
- **Best practices for content recruitment**
 - o What does the literature say?
 - o What can be learned from the repository arXiv's¹ success?
 - o What do the empirics indicate?
- **Co-existence between different types of digital repositories and publishers**
 - o Is a co-existence between digital repositories and publishers/journals viable?
 - o Is a co-existence between different types of digital repositories (institutional and subject repositories) viable, or is just resulting in a redundancy of repositories?
 - If viable, what would be the most natural structure of co-existence?
- **Digital Repositories and Open Access**
 - o Are digital repositories the mean for achieving Open Access?
 - o What type of content do we want in them (peer reviewed or gray)?
 - o Can they manage to get the content?

The empirical part of the thesis is focused towards electronic theses and dissertation (ETD) repositories. What content recruitment strategies that exists at ETD repositories around the world is assessed. Additionally, are the users of the CERN-Theses Database investigated, with the aim of understanding their motives for self-archiving their theses, thinking that this can shed light upon successful recruitment strategies.

The thesis begins with a theory review of the issues pointed out as the focus of attention for this thesis. Next, the reader will be presented the methodology applied during the work, before empirical findings will be presented. Then the questions of the thesis will be discussed in scope of the theory review and empirical findings, before the final and concluding chapter gathers the investigated threads and aims at pointing out best practices for content recruitment to digital repositories, as well as identifying issues suited for further research. Technical issues like choice of software and implementation are regarded as already well-discussed and developed matters, and will not be discussed in this thesis.

(When speaking of different types of repositories in this thesis, one is only referring to digital repositories for academic institutions or disciplines.)

¹ [Hhttp://arxiv.org/H](http://arxiv.org/H)

Section 1 – Theory Review

A presentation of the literature written about the topic of digital repositories is given in the following section. As a natural start comes a definition of what digital repositories, their purpose and extent are. Repository's history, purpose and livelihood cannot be fully understood without being seen in a context, a chapter is therefore devoted to presenting repositories seen in relation to libraries, publishers and Open Access. This is followed by a chapter which deals with the important issue of content recruitment to digital repositories. The theory section ends by summing up the theory review, and indicating in which direction the thesis will continue.

8 Digital Repositories

8.1 Definition Digital Repository

A digital repository is a web accessible database containing research material from scholars. These databases serve the purpose of enhancing access to academic research output and as a location for preservation of scholarly material. Some digital repositories exist in connection to an institution, whilst others are connected to a community or field of research.²

A more elaborate definition of an institutional repository is that it can be defined as 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 (Lynch 2003). To a growing extent institutional repositories are now also spreading to colleges and research institutes, as well as to more commercial actors (Jones, Andrew and MacColl 2006), but the latter is outside the scope of this paper.

Subject repositories on the other hand include digital papers from authors all around the world covering one or more disciplines (Bailey Jr. 2006), and are just as often referred to as disciplinary repositories. Most subject repositories today are based on material harvested from institutional repositories, but a minority attracts content directly from the authors (Swan 2008). Subject repositories first emerged from communities that had a prior tradition of exchanging preprints even before the internet came about and facilitated it, and where the speed of dissemination contra publishing is an important factor (Kling and McKim 2000). Subject repositories often serve as a parallel publishing of material, in order to ensure earlier and more efficient dissemination (Björk 2004).

Documents relevant to the output of a single research institution are much more diverse than those intended for formal peer review, and Genoni emphasizes that an institutional repository should not merely duplicate the existing channels of formal scholarly publishing. (Genoni 2004) Harnad and Brody is of another opinion, arguing that repositories should only

² [Hhttp://www.library.uiuc.edu/scholcomm/glossary.html#d](http://www.library.uiuc.edu/scholcomm/glossary.html#d)

contain peer reviewed papers in order to avoid drowning in available information and to be able to ensure quality of content (Harnad and Brody 1991).

8.2 Their Purpose

In the report called DRIVER, the product of five European Union founded research projects onto the topic of digital repositories, Alma Swan sums the functions and foci of digital repositories in the following manner (Swan 2008);

- To open up and offer the outputs of the institution or community to the world
- To impact on and influence developments by maximizing the visibility of the outputs and providing the greatest possible chance of enhanced impact as a result
- To showcase and sell the institution to interested institutions – prospective staff, prospective students and other stakeholders
- To collect and curate digital outputs(or inputs, in the case of special collections)
- To manage and measure research and teaching activities
- To provide and promote a workspace for work-in-progress, and for collaborative or large scale research projects
- To facilitate and further the development and sharing of digital teaching materials and aids
- To support and sustain student endeavors, including providing access to theses and dissertations and providing a location for the development of e-portfolios

These points make not division between subject and institutional repositories, but one can see that some of them are more fitting for the one type than the other.

8.3 The Extent of Digital Repositories

The extent of institutional repositories in the academic sector is varying from country to country. A survey on the deployment of academic institutional repositories as of mid 2005, showed that countries like Norway, the Netherlands and Germany had a deployment rate of a 100%, whilst Finland could only show to a meager 5%. Not surprisingly the same country wise variation is to find amongst the extent of national programs promoting the deployment of institutional repositories as well as best practices and standards. (Van Westrienen and Lynch 2005)

The Open Directory of Open Access Repositories³ now has over 1300 hundred (academic) digital repositories in its directory (or over a 1000 organization with repositories). From these about 200 were added to the directory in 2007 and 300 in 2008, reflecting the rapid spread of repositories around the world. Out of the repositories in the directory about 50% are European and 30 % North American. Several expect that within a fairly short timeframe, every research based institution in Europe will have a repository (Swan 2008).

Several national and international incentives exist to further the extent of digital repositories. Common features for the majority of the institutional repositories related

³ [Hhttp://www.opendoar.org/](http://www.opendoar.org/)

initiatives are that they focus/ have focused on issues like creating nationwide standards for institutional repositories or EDTs, mainly in terms of technical issues, as well as they conducts research projects on success factors for institutional repositories. There are also several consortiums cooperating on jointly creating an institutional repository for their institutions, like DiVA and the White Rose Research Online Portal. DiVA is a repository portal which currently 23 Scandinavian universities are using. This means that research publications and student theses are uploaded at the home intuition, and can be found both at their database using a local interface, as well as at the joint DiVA portal⁴. Similar is the White Rose Research Online portal, which is a joint institutional repository for three UK universities⁵. ARROW is an Australian initiative working to identify the best choice within software for Australian IR's, as well as to create a national finding portal based on harvested metadata from Australian Institutional Repositories⁶.

Other initiatives worth mentioning are Sherpa and The Networked Digital Library of Theses and Dissertations (NDLTD). The latter is an international organization focusing on the promotion of adoption, creation, use, dissemination and preservation of electronic theses and dissertations⁷. Sherpa is the host of the database OpenDOAR which lists all serious Open Access repositories. Sherpa also host the RoMeo database, which serves as useful aid for consulting copyright issues related to repositories and publishers.⁸ Furthermore the organization also has also been involved in activity related to DRIVER⁹ and EThOS¹⁰.

8.4 Interoperability across Repositories

The initiative that can be said to have gained the most momentum when it comes developing and spreading an international worldwide standard for repositories is the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH). The organization develops and promotes interoperability standards which make it possible to harvest/gather Metadata from repositories that are OAI-PMH compliant¹¹. By interoperability one means that Metadata from one repository can be harvested by another repository and that harvesting metadata across several repositories through a joint portal is possible. Within the OAI-PMH system there are two important roles, the data providers and the service providers. The former are the ones guarding the Metadata as well as making it accessible by being OAI-PMH compliant, the service providers can then offer a value added service across all the

⁴ [Hhttp://www.diva-portal.org/smash/aboutdiva.jsf](http://www.diva-portal.org/smash/aboutdiva.jsf)

⁵ <http://eprints.whiterose.ac.uk/docs/information.html>

⁶ [Hhttp://arrow.edu.au/about/](http://arrow.edu.au/about/)

⁷ [Hhttp://www.ndltd.org/](http://www.ndltd.org/)

⁸ [Hhttp://www.sherpa.ac.uk/projects/](http://www.sherpa.ac.uk/projects/)

⁹ Digital Repository Infrastructure Vision for European Research (DRIVER) is an EU funded organization working for creating a good infrastructure between Pan European digital repositories, and which has been involved in several research projects on the topic of repositories. (<http://www.driver-repository.eu/>)

¹⁰ Electronic Theses Online System (EThOS) works for better access to UK theses. (<http://www.ethos.ac.uk/>)

¹¹ [Hhttp://www.openarchives.org/OAI/OAI-organization.php](http://www.openarchives.org/OAI/OAI-organization.php)

repositories, like for example a field specific aggregator like subject repository.¹² Currently nearly 1000 repositories are OAI compliant.¹³

Institutional repositories can be an institutions window out to the world and serve as a good marketing tool, but as Björk points out its potential to have significant global impact is dependent on them being interoperable and combined with field specific aggregators like subject repositories. He compares the value of repositories to that of mobile phones, it is not truly valuable until everyone got and uses them. (Björk 2004)

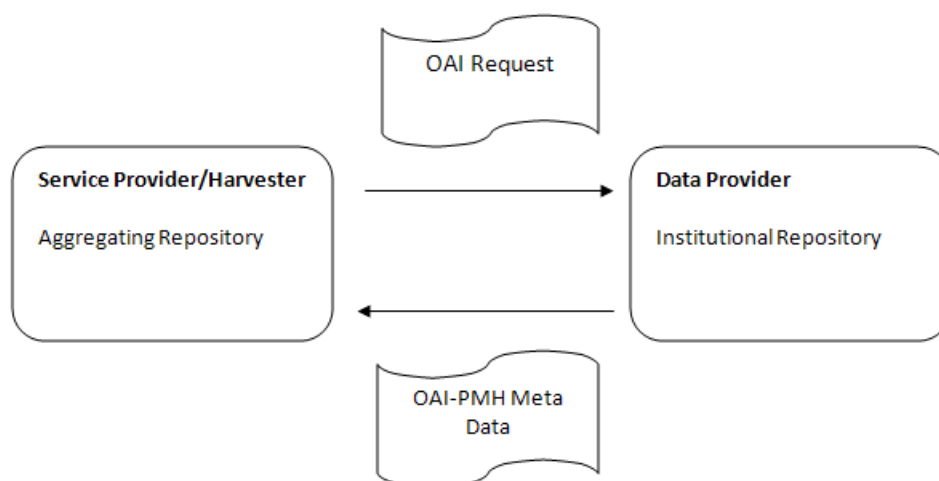


Figure 1 - Repository Interoperability (From Jones, Andrew and MacColl 2006)

8.5 Business Models for Digital Repositories

Repositories are by many seen as the mean for Open Access to all public funded research, an ethical tool and a not for profit concept. But regardless of such noble intentions the repository has to be based on a viable business model if it is to be able to serve its purpose in a sustainable manner.

Swan has identified five different operational models for repositories (Swan 2008);

- Institutionally owned: institutions own and run the business to further their own goals and strategies
- Public bodies sponsor the business for the public good
- The business runs on a community basis, sustained by the communities that they serve
- The business runs on a subscription basis, selling products or services to customers paying cash

¹² [Hhttp://www.oaforum.org/tutorial/index.php](http://www.oaforum.org/tutorial/index.php)

¹³ [Hhttp://www.openarchives.org/Register/BrowseSites](http://www.openarchives.org/Register/BrowseSites)

- The business runs on a commercial basis (other than subscription based): a number of sub-types are covered by this term, for example an advertising model.

Institutional repositories are naturally enough most often institutionally owned, but the other operational models are also represented. Whichever business model is found best suited for a digital repository, it is important that this model is viable, sustainable and adaptable in that specific case (Swan 2008).

In terms of the long term sustainability of a repository it has been implied that institutional repositories are in a better position than a subject repository. Institutional repositories can be regarded as a formal institutional service, and Bailey claims them to be stable and permanent. (Bailey Jr. 2006) Björk argues for the same thing: institutional repositories are by virtue of their host universities and libraries in a better position to ensure long term preservation and operation than subject repositories (Björk 2004). The stability and long term livelihood of a subject repository is dependent on its formal affiliation to an institution or organization and its ongoing interest from authors and readers. (Bailey Jr. 2006)

8.5.1 Are Institutional and Subject Repositories Suited for a Co-Existence?

The existence of the several kinds of repositories raises the questions of the livelihood of them all.

Pinfield argues that repositories should be institutional, as the institutions have the technical and organizational infrastructure and the resources to maintain the repository in the long run. However for this to be reasonable, the institutional repositories have to be interoperable, which can be achieved through being OAI-PMH compliant, and co-existing with field specific aggregators, like a subject repository, which offers effective search services. On these assumptions Pinfield reasons that institutional and subject based repositories are apt for coexistence and complements each other. (Pinfield 2004)

Björk is of the opinion that institutional repositories in the long run can outdo disciplinary ones as they aim at for the same material, and the institutional repositories have the advantage of more viable business model and by good interoperability and search ability can fill the place of the subject based repositories (Björk 2004).

9 Repositories in a Context

The figure underneath illustrates the main stakeholders that make up the environment which digital repositories exist within, the stakeholders presented are based on Crow's explanation of which bodies that are the major stakeholders of digital repositories (Crow 2006). In this chapter digital repositories will be presented seen in the context of libraries, journals and publishers and the Open Access debate, in order to gain a better understanding of the circumstances which under repositories exist today.

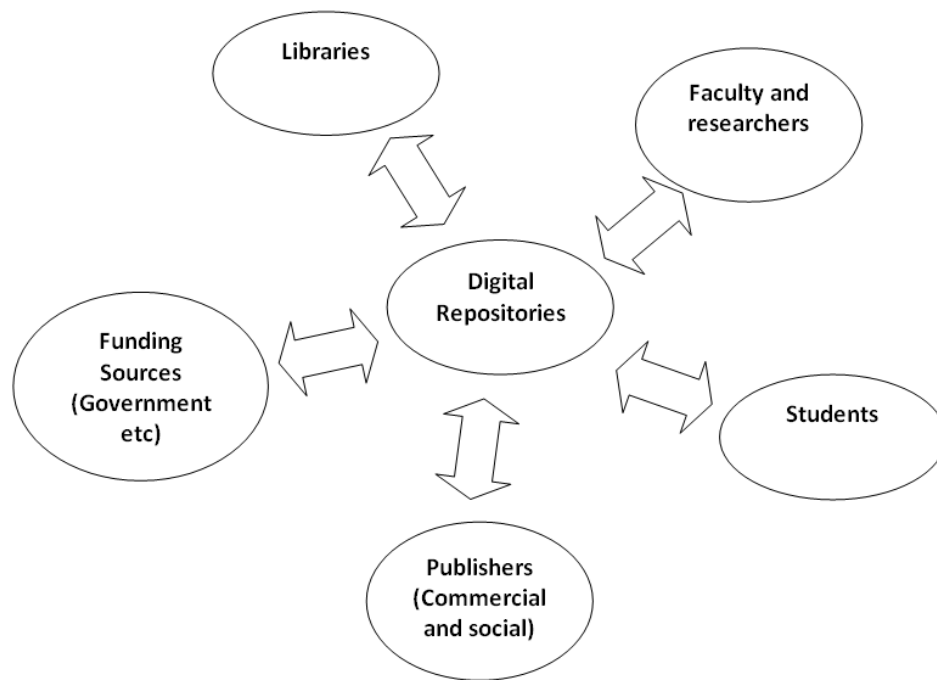


Figure 2 - Digital Repositories and Their Stakeholders

9.1 Libraries and Repositories

Institutional repositories are traditionally associated with a university and its library. The library is typically the actor establishing and maintaining the repository, sometimes in cooperation with the institutions information technology unit. (Bailey Jr. 2006)

“Information and knowledge are expanding in quantity and accessibility” Kofi Annan said back in 1997(Edwards 2001). A statement that by now almost can be referred to as an understatement and rather described as an explosion or revolution within information and the possibilities due to technical developments. Myriads of papers are written about the challenges the libraries and librarians are facing due to the “information revolution” and the technical developments on the ICT side the last couple of decades. (Edwards 2001, Stueart 1988) The mass of information they have to manage these days are of a completely different size than before, an information jungle, and the channels which the information circulates within has changed just as much.

The roles of the library has expanded from the traditional function of collection development, acquiring materials from publishers and other sources to meet the needs of

their users, to now taking a form of publishing role through their work of supporting and building institutional repositories for faculty output (Kreitz and Brooks 2003)

Librarianship is now better defined by words like library science or information science, and information technology is becoming an every increasing important part of the package that libraries have to deal with (Stueart 1988).

Some people claim not to go to the library anymore, what they don't realize then are that the library is coming to them. (Borgman, 2002) A repository can in many ways be such a feature.

9.2 Digital Repositories, Publishers and Journals

There are two opposing philosophical camps among those who work to justify institutional repositories: one that views institutional repositories as competition for traditional publishing, the other that sees institutional repositories as a supplement to traditional publishing (Davis and Connolly 2007).

In order to see whether repositories pose a real threat to publishers and journals an investigation was done on the usage of astrophysics e-prints compared to the published articles. The results showed that as soon as the published version is available, scientists choose the published version over the e-print. Thereof one could draw the conclusion that digital repositories do not pose a threat to the journals and publishers. (Henneken et Al. 2006) This also implies that most scientists have access to journals through their affiliated institutions subscriptions. The investigation was done on usage information from the physics related subject repository arXiv, thus one does not know if the same findings would appear for another field. Henneken et Al. do not only argue that digital repositories poses no threats to the journals, they argue that the repositories increase the visibility of the journals (by the links to the published papers which they often have).

Opposing conclusions stem from an international survey of librarians' preferences with the purpose of shedding light onto whether digital repositories pose a threat to subscription based journals. The survey was sent out to over 25 000 email addresses, about 9000 bounced back leaving the efficient research sample to about 16 000, whereof 424 answered (2,6%). The majority of the answers came from librarians at research institutions or universities in North America and Europe. The statistics of the answers showed a large consistency amongst the respondents, indicating that the rather small but assorted sample can be representative of librarians' preferences. This survey concludes that there is a real threat. The conclusion is made on the basis of the results that showed how librarians would have a strong preference of Open Access materials as opposed to journal subscriptions. A significant number of librarians would be likely to substitute journals subscriptions by digital repositories under the conditions that they found the content reliable, peer-reviewed and up to date. If the content in the digital repository on the other hand was not the final peer reviewed versions, or delayed by an embargo larger than 6 months compared to the journal, the majority of the librarians would prefer the journal subscription. (Beckett and Inger 2007)

9.2.1 The Functions that Constitutes Scholarly Communication

In this subsection the functions of scholarly communication will be explained in order to better understand the roles of the journals and the repositories. Scholarly communication has been defined to consist of the following five functions:

- *Registration*, which makes it possible for the scientist(s) to claim the credit of a finding
- *Certification*, which serves the role of validating the registered findings
- *Awareness*, which serves to make the scholarly community aware of the new claims and findings
- *Archiving*, which has the responsibility for long term preservation of the scholarly output
- *Rewarding*, the rewards given to a scholar is based on metrics reflecting the scholar's performance within the value chain of scholarly communication.

(Roosendaal and Geurts 1997)

Previously all these functions, but archiving, has to a large extent solely been attended to by publishers and their journals. Archiving of paper based formats was traditionally done by libraries around the world. Today, as the common formats are becoming digital, several new actors start taking part in the value chain of scholarly communication. If one uses arXiv as an example one can see how arXiv caters for several of the functions in the value chain of its scholarly community. By submitting ones manuscript to arXiv it has been registered. To some extent arXiv also allows certification, by letting users comment on papers and by a lightweight filtering of papers done by volunteers. Still arXiv's users rely on the more formal certification done by the journals peer review. Finally arXiv also has the function of creating awareness of the research material (by having all its content freely available at the internet, letting search engines index their content and a newsfeed function to interested scholars) and archiving by ensuring preservation through redundancy on their servers. (Van de Sompel et Al. 2004)

As the arXivs example shows, certification and rewarding are the only functions the repository do not cover to a full extent. This fact has been pointed out by several scientists, which have suggested different implications. Certification is mostly based on traditional peer review. A process that delays the time from a manuscript is born to the time it is published, and cannot keep track with the rapid means for dissemination that the internet has brought about, in the same way it could keep pace with paper based dissemination. Ginsparg (2004) suggests that peer review should adapt to developments within scholarly communication by changing and splitting up the process, firstly having a rapid and low cost screening of papers, merely establishing that they are of "to be peer reviewed quality" and leaving for the expert, but speed concerned, user to further judge the content. Secondly having a more stringent peer review process of the papers fulfilling certain criteria, aiding the less expert and time

conscious users as well as being an aid for navigation through the masses of existing papers. (Ginsparg 2004)

9.3 Open Access and Repositories

Open Access became a movement in 2001, after a meeting in Budapest held by the Information Program of the Open Society Institute (Guédon, 2004).

9.3.1 Definition of Open Access

The Open Access movement seeks to give anyone, anywhere at any time unrestricted access to all scholarly research material (Bianco et Al. 2007). The Budapest Open Access Initiative (2002) defines Open Access literature as literature that has:

“ free availability on the public internet, permitting any user to read, download, copy, distribute, print, search, or link to the full texts of these articles, crawl them for indexing, pass them as data to software, or use them for any other lawful purpose, without financial, legal or technical barriers other than those inseparable from gaining access to the internet itself. The only constraint in reproduction and distribution, and the only role for copyright in this domain, should be to give authors control over the integrity of their work and the right to be properly acknowledged and cited”. (Budapest Open Access Initiative 2002)

9.3.2 The Open Access Movement – Why?

Research is mainly funded by public money (Guédon 2004). Thus wanting the public to have Open Access to the results is not an unreasonable argument. But what drove the Open Access to where it is today? Several authors mention “the Scholarly Communication Crisis” or “the Scholarly Serial Crisis” as a main reason (Chan 2004, Swan 2006). The term is referring to a two faced problem. The first part is the *affordability problem*; subscription fees (particularly for science and medical fields) have by far outgrown inflation rate the previous two decades (a monitoring of the journal prices in the UK showed a increase in the subscription fees from 24% to 97%, over the time span from 2000 to 2004, this compared to an annually inflation rate of 2,5% (Swan 2006)). So in spite of growing library budgets, the number of subscription is decreasing. The second part of the crisis consists of the *accessibility problem*; subscriptions are bundled in databases controlled by large commercial actors. These gives very restrictive access to who can access the database and when. As most such subscriptions only come in an electronic version, it leads to the user not being able to access to older “material already subscribed for” if the user ends his subscription. In other words does not a subscription at one point in time, ensure the long term access to the material. (Chan 2004)

This is evidently not a desirable situation for the scholarly community. On one hand they meet difficulties in accessing the information they need due to restricted access, and on the other hand the restricted access can also be seen as a barrier to the dissemination of one ones own results which is necessary in order to have impact and being visible. For third

world countries restricted access and expensive subscriptions are even more critical (Mark and Shearer 2006)

9.3.3 Green and Golden Open Access

There are many different alternatives for “organizing” Open Access. Harnad, known as an eager proponent of Open Access publishing, introduced the terms green and golden Open Access, depending on the type of “openness”. (King and Alvarado-Albertorio, 2008) Golden Open Access is referring to publishers/journals that rely on a business model that do not depend on subscriptions and restricting access. Some golden Open Access journals survive on funding from non profit organization or host institutions. Such journals are more the exception; with the rule being that the journals needs some kind of revenue to survive. One way of achieving (golden) Open Access is the Author Pays model where the author pays to have its material published in the journal (which generally means that his affiliated institution pays an “article processing charge”) and the journal subsequently gains its revenues in this way instead of from subscriptions. (King and Alvarado-Albertorio 2008)

Green Open Access on the other side is where the repositories come in. Green Open Access is literature in freely accessible repositories (either institutional or subject based). Institutional repositories and/or subject repositories in this way have the potential to play a significant role in the quest for Open Access. Many argue for the co existence of both the green and golden options. (King and Alvarado-Albertorio 2008) Green Open Access to literature that also appears in journals, are dependent on that the journal allow self-archiving of the material they publish (Guédon 2004).

9.3.4 Proponents and Extent

A survey on the amount of Open Access literature that exists indicates that “green” publishers control around 85% of the scientific titles published in the world (Guédon 2004). Something which means that restrictions in terms of self-archiving and copyright cannot be largely blamed for the difficulties many repository administrators have filling their repository. In comparison Open Access journals is found to cover only about 5 % of the scientific literature (in number of articles) (Guédon 2004). From these numbers one could assume that green Open Access should be much more viable and efficient than golden Open Access, but it has to be emphasized that even though 85% of scientific literature is allowed to self-archive, this is far from being done. Green Open Access cannot alone ensure accessibility; it is reliant on well functioning repositories.

In which form one will see Open Access in the future is not certain, but that there will be changes in direction of more freely available scientific literature is to be expected as the Open Access movement gains momentum. Over 5000 individuals and 320 organizations have signed the Budapest Open Access Initiative that came about in 2001.¹⁴ “The Berlin Declaration on Open Access to Sciences and Humanities” which was the result of a 3 day

¹⁴ [Hhttp://www.soros.org/openaccess/view.cfm](http://www.soros.org/openaccess/view.cfm)

long conference in 2003, is signed by more than 260 institutions, which by their signature promise to work for and support Open Access¹⁵. This is only a couple of many statements supporting the Open Access movement¹⁶. Additionally several other institutions work for and support the idea of Open Access. CERN has long been one of the Open Access proponents.

9.3.5 CERN and Open Access

Open Access at CERN (the European Organization for Nuclear research) has a long history, the CERN Convention of 1953 states: "...the results of its experimental and theoretical work shall be published or otherwise made generally available"¹⁷. The collaboration "Sponsoring Consortium for Open Access Publishing 3" (SCOAP³)¹⁸ in high energy physics was initiated by the previous Director-General, Robert Aymar, at CERN and is now run by CERN's library department. The SCOAP³ collaboration has developed a business model for the HEP Open Access publishing based on the idea that all countries that published HEP papers should contribute to the cost of publishing based on their percentage of the total HEP papers published. The SCOAP³ collaboration will collect and manage the monetary contribution of all the participating countries. In order to allocate the budget amongst the journals that publishes HEP papers, there will be performed a tendering process. The allocation will be based on the number of articles that the journal publishes, the cost per article that they state in their tenders and their scientific quality. This is a unique and innovative approach to Open Access, which if successful will have large influence on the whole publishing industry related to HEP.

9.3.6 What Is Not to Like?

Harnad questions this organizing of golden Open Access in terms of quality insurance and the fact that such a model means that the journal increases its revenue by letting more authors publish, and therefore might lower its standards when peer reviewing. (Harnad 2008) Other argue that journals can only attract submission, if they appear as a journal of quality, and therefore it would not be inclined to lower its quality standard, in spite of the short term economical incentive (Cockerill 2006). Several also question the quality of content that are to be found in repositories, which often is the archive of an institution or

¹⁵ [Hhttp://oa.mpg.de/openaccess-berlin/berlindeclaration.html](http://oa.mpg.de/openaccess-berlin/berlindeclaration.html)

¹⁶ OATS – Open Access Team Scotland ([Hhttp://scurl.ac.uk/WG/OATS/institutionalsupport.htm](http://scurl.ac.uk/WG/OATS/institutionalsupport.htm))
Budapest Open Access Initiative ([Hhttp://www.soros.org/openaccess/read.shtml](http://www.soros.org/openaccess/read.shtml))

Bethesda Statement on Open Access Publishing
([Hhttp://www.earlham.edu/~peters/fos/bethesda.htm](http://www.earlham.edu/~peters/fos/bethesda.htm))

Welcome Trust Position Statement on Open Access ([Hhttp://www.wellcome.ac.uk/About-us/Policy/Policy-and-position-statements/WTD002766.htm](http://www.wellcome.ac.uk/About-us/Policy/Policy-and-position-statements/WTD002766.htm))

OECD Declaration on Access to Research Data from Public Funding ([Hhttp://www.oecd.org/H](http://www.oecd.org/H))

Institutional Self-Archiving Policy Commitment - Registry of Open Access Repository Material Archiving Policies ([Hhttp://www.eprints.org/openaccess/policysignup/H](http://www.eprints.org/openaccess/policysignup/H))

World Summit on the Information Society (WSIS) ([Hhttp://www.itu.int/ws/is/index.html](http://www.itu.int/ws/is/index.html))

¹⁷ <http://open-access.web.cern.ch/Open-Access/pp.html>

¹⁸ [Hhttp://scoap3.org/](http://scoap3.org/)

communities research output as whole, not only peer reviewed papers that has gone through the quality control associated with publishing in a recognized journal. Typically they support author deposit and metadata creation, and keyword and field searches. The author depositing of material makes it possible to run a low cost repository, but on the problematic side, the administrator has little not enough resources to do any quality reassurance of the content (Björk 2004).

Guédon points out that there is an evident inability to convince 100% of authors to self-archive 100% of their articles, and that this is a great barrier to successful green Open Access (Guédon 2006). Furthermore he is of the opinion currently journals can be seen as one of the greatest threats or competitors to green Open Access, due to the fact that most researchers can access the literature they need through their affiliated institutions journal subscriptions, and that these journals are able to offer more value to the reader than repositories because the former currently have better search ability and user interfaces (Guédon 2004).

10 Content Recruitment

Content recruitment has earned its own chapter from the simple reason that it by the author is regarded as the single most important factor one should successfully manage in order to have a healthy repository. (That is when ignoring technical issues related to software etc, which is outside the scope of this paper.) The majority of the literature on the topic does not explicitly say what is regarded as a success for a repository, but it seems like the unison perceived opinion is that a good coverage of “planned” content is equivalent with success.

A poor coverage of the material a repository aims to contain makes it of significantly less value for the individuals using it to search for information. Furthermore it might inhibit authors to upload their work to the repository, leading to a vicious and negative circle for the repository (Davis and Connolly 2007). If one leaves the repositories that get their content merely from harvesting other sources out of account, faculty or author participation becomes alpha and omega. In spite of the fact that waste resources around the world are being put towards the creation of digital repositories and running them, many of them remain empty shelves that no one pays any attention to. The phenomenon is worldwide, even with a variety of creative ideas and promotional activities; faculty uptake has been reluctant where voluntary compliance is needed (Mark and Shearer 2006). Two surveys performed in 2005 on institutional repositories in respectively the United States and Europe both reported about meager participating from faculty when it comes to uploading of content (Lynch and Lippincott 2005, van Westrienen and Lynch 2005).

Due to these reasons research has been done the previous years on the process of content recruitment to digital repositories. In the follow sections common strategies for content recruitment, reasons found for failure of content recruitment as well as success factors will be presented. The two latter will naturally enough overlap to some extent. Most of the research on this field has been directed towards university affiliated institutional repositories. Only a minority of the research has been especially directed towards ETD repositories, but as the majority of ETD repositories exist in relation to an institutional repository, it seems reasonable to assume that the research findings can be transferred from one to the other to some extent.

10.1 Strategies for Content Recruitment

From the research done by the Canadian Association of Research Libraries (CARL) in 2006, on what content recruitment strategies that exist for institutional repositories, six categories of content recruitment strategies were found; general promotional activities, depositing services, content harvesting, researcher bibliographies, usage and citation information and university policies. Under the findings within each category will be listed, most points are self explanatory and will not be elaborated on (Mark and Shearer 2006);

General Promotional Activities

- Passing out brochures
- Conducting presentations for faculty committees

- Publishing articles in the library/school newspaper
- Formally launching

Depositing Services

- Copyright checking activities
- Metadata assignment
- Library depositing of content

Content Harvesting

Having the repository administrators collect material themselves, might be a good way of initially seeding the repository, and is also done by some repositories on a long term basis.

Harvesting can be done in the following ways;

- Gathering material from the institutions researchers web pages
- Depositing articles from Open Access journals and other journals that allow self-archiving

Researcher Bibliographies

- Personal web pages for the researcher in the repositories, with information about the researcher and links to his publications within the repository

Usage/Citation Information

- Providing usage information can attract authors (how many times has the article been downloaded, what are the ten most downloaded articles etc)
- Demonstrate how papers in repositories (Open Access papers) are more often download, thus contributing to improving the visibility and impact of the researches work.

University Policies

- Policies that encourages self-archiving
- Mandatory policies (recommended, can be highly effective but can also be hard to implement)

As the CARE research paper admitted, several of the institutional repositories and therefore also their content recruitment strategies are to new to be proved as having long term efficiency.

10.2 Reasons for Poor Faulty Uptake

The lack of awareness of the actual existence of the respective institutional repository is found as a main reason for poor faculty uptake (Swan 2004, Swan and Brown 2005), as well as a lack of mandatory policies for the deposition of work (van Westrienen and Lynch 2005).

Other reasons revealed are the concerns for copyright and intellectual property issues. The authors are not sure whether they might be violating copyright agreements that they have signed with their publishers, or whether posting a document in a repository will be considered as a prior publication. (Mark and Shearer 2006, van Westrienen and Lynch 2005).

Lack of perceived incentives and benefits by researchers of self-archiving their work is detected as yet another issue for poor faculty uptake. Self-archiving of one's work comes as an additional task which cuts into one's research and writing time, on top of an already full workload and is thus not prioritized. (Mark and Shearer 2006, Foster and Gibbson 2005) Pinfield points out that all existing reward-mechanisms are based on the current publishing system, not on self-archiving (Pinfield 2004).

An occurring problem when repository administrators perform the general promotion of the repository is thought to be that the librarians (i.e. the administrators) do not speak the same language as the faculty, and that the communicated attributes regarded as beneficial by the librarians might not mean anything to the faculty. (Promoting persistent URLs and metadata schemes might sound endearing for a librarian, whilst it means nothing to a researcher.) (Mark and Shearer 2006)

In some cases a misfit between the benefits and services of an institutional repository and the needs and desires of the faculty can be an attributing factor for poor faculty uptake. In other words there is often a need for gaining a better understanding of the work practices of the faculty in order to both promote and adjust the institutional repository to a better fit with faculty needs. The IR must support the faculties existing ways of working. (Foster and Gibbson 2005) Foster and Gibson also speculates in whether the name institutional repository is unappealing to the faculty/authors, and that a name like 'personal digital repository' might communicate the benefit for the individual contributor in a better way.

Perceived time required and possible technical difficulties are an important barrier to self-archiving. (Swan 2005) Additionally some researchers might regard the institutional repository as an island as opposed to a subject repository that might be perceived as further reaching (Davis and Connolly 2007).

Cultural differences (tradition for publishing and sharing of work etc.) between disciplines are found to be a reason for a varying success of content recruitment amongst discipline. (Kling and McKim 2000) Disciplinary related investigations that have mapped faculty participation in depositing of e-prints into repositories found that physicists and astronomers had the highest participation, followed by mathematicians and computer scientists, engineers, cognitive scientists and psychologists, and biological scientists. Chemists reported no participation. (Lawal 2002)

An investigation on self-archiving in the UK in 2003 found a correlation between willingness to self-archiving and the existence of subject repositories. Within the fields that had well established subject repositories self-archiving on personal or departmental web pages was common, indicating that academics within these disciplines might even be willing to duplicate their effort by self-archiving to both the subject repository in addition to their departmental/personal webpage. The exception was however amongst physicists which are the discipline said to have the most successful repository. It was suggested that this was due to arXiv becoming so successful in capturing and making persistently available a large

proportion of the literature on the field, that it amongst physicist has become the natural choice for self-archiving their material. (Andrew 2003)

Table 1 – Summary of Reasons for Poor Faculty Uptake

Summary of Reasons for poor Faculty Uptake
Lack of awareness of repository
Lack of mandatory policies for deposition
Doubts in terms of copyright
Lack of perceived incentives
Scholarly reward mechanisms are based on traditional publishing
Wrong type of promotion
Misfit with faculty needs
Perceived time required
Disciplinary differences regarding traditions for dissemination of work

10.3 Success Factors

As mentioned earlier, most repositories are so new that it is yet hard to tell which content recruitment strategies that are efficient in the long run, but there is one factor which repeatedly has pointed itself out as the most efficient strategy, namely institutional mandatory policies for self-archiving (Mark and Shearer 2006). Not surprisingly the same has been suggested in the context of ETD repositories. Mandatory policies for uploading EDT's is found to be highly efficient, and thus strongly recommendable (Sale 2006, McMillan 1999). By comparing the theses available in Australia's coordinated ETD gateway to the different institutions repositories throughout Australia (the Australian Digital These Program), one saw that where no mandatory deposit policies exists, coverage was as low as 12%, whilst where mandatory policies existed, deposit rates are close to a 100%. (Sale 2006)

Other aspects found important to take notice of is that each discipline has a normative culture, largely defined by their rewards systems and traditions. If the goal of institutional repositories is to capture and preserve the scholarship of one's faculty, institutional repositories will need to address this cultural diversity. (Davis and Connolly 2007). For faculty the most important criteria are that other people find, use and cite their work. (Foster and Gibbson 2005) And such advantages are thus the ones that should be promoted. Furthermore Foster and Gibbson found the faculty wants the possibility for storing work in progress and document versioning. (Foster and Gibbson 2005)

10.4 ArXiv – An Example of a Successful Digital Repository

When success factors and digital repositories are to be discussed, the subject repository arXiv is seldom failed to mention. ArXiv is a pioneer within its breed and plays an important role within the community it serves. Due to its success, importance, presence in the repository related literature and connection to the main discipline at CERN it is only natural that arXiv gets its dedicated place in this paper.

10.4.1 ArXiv's Background

ArXiv first saw daylight in 1991, when it was founded by Paul Ginsparg at the Los Alamos National Laboratory in New Mexico, and it is now hosted at Cornell University in New York, partly funded by the National Science Foundation. Originally arXiv contained high energy physics (HEP) related e-prints merely, but now the repository has grown to become the reference repository for several other fields, amongst the major ones mathematics and computer science. Little services that could represent extra value added to arXiv exist; it is mostly a plain repository, which has got little emphasis on sophisticated search mechanisms or offerings of bibliographic information. (Gentil-Beccot et Al. 2008) Today the repository carries more than 530 000 e-prints. Their usage statistics shows a steady rise in monthly submissions since its birth, the number now being more than 5000 monthly submission.¹⁹

10.4.2 What Are the Reasons for arXiv's Success?

When the cause of arXiv's birth and success is explained in the literature it is mostly accredited to long lasting traditions for information sharing in the area of HEP, as well as the fact that the participants in field has been pioneers when it comes to scholarly communication. The tradition of sharing preprints amongst physicists dates back to the 1970s. Back then preprints were sent to paper based pre print clearing houses, which redistributed the papers to interested researches (Kreitz et Al. 1996). This habit of distributing the pre prints evolved with the advancements within technology, going from paper prints to the usage of emails and file transfer protocols (FTP) (Kreitz et Al. 1996), and later repositories like arXiv. Other pioneering innovations within scholarly communication and information management that has sprung out from the community of HEP are by example the internet²⁰ and the very first database for grey literature to go online²¹ as well as the first and archetypal digital repository, namely arXiv (Gentil-Beccot et Al. 2008).

As high energy physicists are regarded as the pioneers when it comes to repositories and scholarly communication, several has argued that other disciplines will follow the field of HEP little by little, with the late adopters at the very end, something which is a traditional view of how new technologies disseminate. If this is the case one should think that the model of arXiv and its success should be transferable to other fields. (Kling and McKim 2000) However others argue that this traditional way of thinking when it comes to how a technology disperses, will not apply for the adoption of digital repositories throughout the academic world. This because prior to arXiv's existence high energy physicists already had the habit of disseminating their results before publication by exchanging their preprints, thus leading to such a successful uptake of arXiv's, but that this cannot necessarily be expected to

¹⁹ http://arxiv.org/show_monthly_submissions

²⁰ The World Wide Web was invented by Tim Burners-Lee at CERN in 1989.
(<http://public.web.cern.ch/public/en/About/WebStory-en.html>)

²¹ SPIRES, a database for HEP literature run by Stanford Linear Accelerator Centre (SLAC) since 1960, was the first database to go online and also had the first web site in North America.
(<http://www.slac.stanford.edu/spires/about/>)

apply for other academic fields.(Kreitz et Al. 1996, Kling and McKim 2000). These authors believe that adoption of subject repositories within other fields will be dependent on their own traditions. Insinuating that some fields might not being very likely to ever adopt a subject repository in a successful way. (Kling and McKim 2000)

The founder of arXiv himself states that the reasons for why arXiv has become so important to its users are obvious. It is because arXiv is the tool for instant scholarly communication, by which waiting months for results to be published can be avoided. Furthermore is serves as a comprehensive archival aggregator, where a high energy physicists can know that if a certain paper exists, it is to be found there. (Ginsparg 2004)

Due to arXiv's size and coverage of a certain field, it might by now have become a physicist natural choice of location to self-archive, as opposed to self-archiving in some affiliated institution's repository. (Andrew 2008)

11 Summary of Theory Review and Further Work

In the previous chapters a presentation has been given of the literature concerning the topics and questions that are scoping this thesis. One has learned a great deal about what digital repositories are, the purposes they serve and about the potential they represent in an Open Access setting. But from the review one can also spot some gaps in the literature. In terms of institutional and subject repositories the literature has been rather generalizing when it comes to what their purposes are. Therefore their respective purpose will be discussed in order to emphasize possible differences, with the idea that this would also bring insight to the feasibility of a co-existence between different types of repositories and journals. The prospect of co-existence between different entities in the value chain of scholarly communication has been touched upon in the repository related literature, and this topic will be sought developed further by a discussion. Furthermore content recruitment is one of the main focuses of this thesis. The literature has spoken of many different aids for successfully recruiting content, with the tendency of focusing towards technical aspects which might neglect many other important factors. From the “soft issues” one has learned about through the theory review, it is evident how the knowledge about which measures (besides mandatory policies) for content recruitment that works is limited, and will hence be sought contributed to by the thesis’ empirical investigation.

In the following section the author will discuss the topics raised in the outset of the thesis in light of the existing literature, believing that new questions can be gained insight to by using existing findings and literature if one rearranges it and considers it from new angles. Thus for the majority of the questions the investigation is a merely theoretic exercise. These questions and discussions are viewed as important in the thesis, and as matters of great influence to the livelihood of digital repositories. They are therefore accredited with as much attention as the more concrete issue that could be expanded on by an empirical investigation. The question of best practices for successful content recruitment is accredited with an empirical investigation focused at learning more about existing strategies for content recruitment and what actually motivates a researcher to submit a paper to a digital repository.

Section 2 – Method, Empirical Investigation and Discussion

12 Methodology

12.1 The Background for the Thesis

CERN is one of the largest research institutes of the world and its library, has amongst others, the non trivial task of preserving, disseminating and making accessible (to everyone, everywhere) the research output from no less than 2500 staff and over 8000 visiting scientists. CERNs institutional repository, CERN Document Server (CDS), is their tool for doing so. Further on CERNs main discipline, high energy physics/particle physics are associated with the subject repository, ArXiv, which never fails to be mentioned in the context of successful repositories. This thesis is colored by the fact that the author has been a technical student at the CERN Library during the work with the thesis and thus was inspired by this environment as well as sought to benefit from being situated at an international research based institution who's main discipline has been at the forefront of the development and adoption of digital repositories.

The author of this thesis has no background within computer science, and consequently has not intended the thesis to contribute to the knowledge of technical matters concerning digital repositories. The author is of the opinion that a large proportion of the problems digital repositories are facing today, origins from that the non technical challenges of how to get authors and organizations to use a digital repository has been neglected, and that a merely technical focus when creating a repository or investigating them, might be misjudging what are the most important challenges. From these reasons the author choose to focus on the “soft” issues related to digital repositories, with the aim of contributing to the knowledge of best practices for getting content into them. Additionally the thesis looks into digital repositories in the context of Open Access and publishers, as these matters are by the author considered as essential for marking out the further course of digital repositories and their existence.

12.2 Literature

12.2.1 Searching for Literature

The majority of the literature was found by searching for papers using search keys like institutional repository, digital repository, community repository, subject repository and ETD repository. The initial searches were mostly performed at Google Scholar²², which served to find titles and where they were published. Thereafter the specific papers were gathered from online journals like ScienceDirect²³ and SpringerLink²⁴ through NTNUs subscriptions. Reference list and links to “related articles” were used to find additional literature. (The

²² <http://scholar.google.ch/>

²³ <http://www.sciencedirect.com/>

²⁴ <http://www.springerlink.com/home/main.mpx>

author using Google Scholar as her main search engine (and not a “more academic” one), whereupon not being able to directly access much of the papers, when one could be fooled to think that if any, these are the papers to be found Open Access due to the purpose of the digital repository itself, might be seen as somewhat of paradox.)

12.2.2 Focus of Existing Repository Related Literature

Quite an amount of literature exists on the topic of digital repositories. At the thesis point of departure the author assumed that the main focus of research and literature concerning digital repositories had been related to technical matters. Literature search and review confirmed this assumption. The majority of existing literature where found to be along these lines:

- Case specific descriptions about the process of creating a repository at a certain institution, focusing on how soft ware was chosen and implemented
- Results from national (or multi institutional cooperation) projects carried out in order to determine the best practices for institutional repositories, in terms of type of soft ware and technical standards for file formats etc
- Technical articles describing certain repository soft ware types (E-prints, DiVA, D-Space etc), or technicalities concerning interoperability amongst repositories.

Additionally there exist literature describing what a digital repository is on a conceptual level, their purpose and related benefits for the users (authors and readers) and the host institution (if existing). Furthermore there has been some research on reasons for poor faculty adoption and what can be pointed out as successful content recruitment strategies. Other digital repository literature was found “hidden” under titles and topics mainly focused on areas like the Open Access debate and libraries.

12.3 Empirics

12.3.1 Aim of Empirical Investigation

The aim of the empirical investigation is to “test” and contribute to prior literature on strategies and success factors/best practices for content recruitment to digital repositories. On a more non theoretical level the empirical investigation seeks to point out practical improvements for the CERN-Theses Database. Strategies, practices and experiences from seven ETD repositories around the world, in addition to the CERN-Theses Database, are investigated as well as the motives of the users of the CERN-These Database for submitting theses to the repository were assessed.

12.3.2 Choice of Data Sample

To narrow down the scope of the empirical investigation, it was decided to focus on repositories for electronic theses and dissertations. Theses and dissertations differ from other research related papers in the way that they are subject to a certification process which is quite different from regular publication. Therefore theses and dissertations are

found across a much broader range of subjects than what are covered by the few successful subject based repositories (Warner 2003). In that way ETDs represent a unique population of research material to investigate, and they are also the type of research one can expect to most easily mandate the uploading of, by making it a criterion for graduation. As the author of the thesis was situated at CERN when working on the thesis, the idea was always to benefit from the close proximity to the research object by collecting parts of the empiricals from one of the repositories at CERN. CERN has got an institutional repository (CDS) and an ETD repository (CERN-Theses Database), which both could serve as good objects of investigation. The choice of looking at the CERN-Theses Database is grounded in the idea that ETD repositories are more similar to each other and come in fairly structured formal forms, and thus represent a more lucid research object than a mixture of different types of digital repositories would, and therefore is a more suitable research object within the given time for the project.

Other ETD Repositories to investigate in addition to the CERN-Theses Database were chosen by searching the Open Directory of Open Access Repositories²⁵, with the aim of finding a wide selection of ETD repositories scattered around the world, which had either a significant amount of ETDs or were fairly new. Their websites were also consulted in order to ensure that they were currently “alive” and not some old ETD project that had stagnated. In addition to the CERN-Theses Database, seven ETD repositories were investigated. Users of the CERN-Theses Database were randomly chosen from the ones that had submitted theses in 2008 or 2009. In this context users are thought of as the person submitting a thesis to the database. Users are students, supervisors and administrative persons in different departments at CERN. One supervisor, four students and three representatives from different secretariats at CERN was interviewed. At the beginning of the investigation the author did not know that some secretariats at CERN dealt with submission of theses to the theses database. This came to the author’s knowledge through studying the names of those submitting theses to database. When the author saw that one secretariat frequently uploaded thesis, the author decided to investigate two additional secretariats to learn what routines they had on the matter.

12.3.3 Choice of Empirical Method of Investigation

It was decided to perform the investigation of the ETD repositories per an email questionnaire. This was viewed as the best alternative to interviews (by phone) which would have been too time consuming for the given project. Phone interviews would probably have given more detailed results, but it would not have been possible to cover as many repositories as when relying on email. Answers from between 5 and 10 repositories was regarded as suitable for the project, representing a feasible amount of information to process and sufficient respondents to expect to get a feeling of the assorted experiences and practices existing at ETD repositories, hopefully also coming across some new insights. The same investigation of the CERN-Theses Database was done by interviewing the head of the CERN Library, Jens Vigen, which due to geographical proximity was naturally to interview face to face rather than by email.

²⁵ [Hhttp://www.openoar.org/](http://www.openoar.org/)

The users for the CERN-Theses Database were chosen to interview face to face. Alternatively it was considered to investigate them by sending out a questionnaire to a significant amount of users, but it was decided that due to the nature of the topic of the investigation, motives for submission, it was easier to uncover the primary motive of the user in a face to face interview. A finding that was thought to be more interesting, than a numerical overview of different motives for the users. It was also reasoned that if gathering the empirics by questionnaire the results might be strongly dependent on being able to pose “the right” questions and suggest the “right alternatives” and thus limiting new insights that could be more easily discovered during a dynamic interview.

12.3.4 The Empirical Investigation Process

The first part of the empirical investigation was to create a suitable email template for a questionnaire addressed at ETD repositories, in order to learn about their practices, strategies and experiences, inspired by what theory had said about this topic earlier on. Fifteen repositories were chosen and contacted by email to ask whether they were willing to contribute to the investigation by answering an email interview/questionnaire. Eleven were positive to the request, but in the end seven sent their answers. The administrators of the ETD repositories were sent the questionnaire in pdf-format, which they could answers directly in the pdf and email back. Some questions were aimed at getting plain background information like the size of the repository (number of theses and dissertations) and what year the repository was created, but also to “warm up” the respondent. The remaining questions were a mixture of questions with suggested alternatives as well as “open” questions. The thought behind this approach was that the eager respondent could share their experiences on the open questions, possibly pointing out aspect new to the investigator, and thus not exist amongst the given alternatives. The less eager respondent might not be inclined to answer the open questions, and rather just want to tick of alternatives found representative by the respondent. To avoid misunderstandings, answers that seemed unclear to the investigator were followed up by email. The questionnaire can be found in appendix 17.3, as well as a list of the responding repositories (appendix 17.2). Next the same investigation of CERN practices, strategies and experiences for their ETD database were conducted.

As the final part of the investigation users of the CERN-Theses Database where interviewed with the aim of learning about what encouraged them to upload their thesis. A varying number of agents from the different user groups were interviewed, to understand what were the drivers motivating them to upload a thesis to the database, but also how they came to know the CERN-These Database and what their impression of the usage/knowledge of it was in their respective environments. It became evident that every interview had to be adapted to each respondent’s level of knowledge of repositories and Open Access. Some had little opinions to share and it was rather coincidental how they had come across the CERN-Theses Database. Others of the respondents were more conscious users and thus also had more opinions on the topic. Naturally enough the interview also had to be adapted to whether the interviewee was a student, a supervisor or someone from the administrative

staff at CERN. The interview template can be seen in the appendix 17.4, but the interviews tended to drift from this template due often stumbling across new findings concerning the working habits and repositories at CERN that were interesting to follow up on. As can be seen from the interview template, there were also some questions related to technical issues like how the interviewee found the submission applications. This information was merely passed on to the library staff at CERN, and is not regarded as a part of this thesis.

12.3.5 Validity and Critique of the Results

As briefly mentioned earlier, these results have no numerical significance, and one cannot know how applicable best practices found in one situation are in another situation. The reader should therefore be aware that generalizations are made on the basis of the experiences from a small and varied population, and that the result might not be relevant in all contexts. What the results do are to uncover new insight to a fairly new area, and can give repository creators or researches ideas for along which lines to proceed.

13 Empirical Investigation

13.1 Research Questions

The primary aim of the empirical investigation was to learn more about best practices for content recruitment. The email questionnaire that went out to the ETD repositories was thus structured to learn about the points below:

- Strategies for content recruitment at ETD repositories
- Experiences about which strategies that were successful for recruiting content to ETD repositories

(In addition to the CERN-Theses Database seven ETD repositories contributed to the investigation.)

The interviews of the users of the CERN-These Database were designed to uncover the following question:

- What are the motives of the users for uploading their thesis?

(Representatives from three different main secretariats at CERN (out of a total of six main secretariats), one supervisor and four students were interviewed about their motives for submitting theses to the CERN-Theses Database.)

13.2 Findings from Investigation of ETD Repositories

In the following section the main findings from the investigation of the ETD Repositories will be listed.

13.2.1 Mandatory Policies

Three of the institutions had mandatory policies for submission of E-theses, additional one institution had a mandatory policy for some of its faculties, whilst the three remaining counted on voluntarily uploading. Only one of the institutions did not have any plans of making it mandatory.

13.2.2 Promotional Activities and Incentives

The most common promotional activity (five institutions stated to do this) was to conduct work shops or presentations of how the submission procedure of the E-theses works, and of what benefits that are involved when making your thesis available online. Such presentations were given both to students and faculty, the students being either new or close to graduation.

Further on several of the institutions stated to have passed out brochures, as well as information packages to new students, as a promotional activity. Some had also published articles in the library newsletter describing the submission process and related benefits.

Three of the institutions explained that their students did not have to hand in paper copies (or at least fewer copies), if they chose to submit their theses electronically.

Two institutions stated to have an informative web page which served to promote the submission of ETDs, but one can assume that this is something the majority has got, in spite of it being failed to mention (as it is probably to be regarded as a rather passive way of promotion).

13.2.3 Content Harvesting

Four of the institutions digitized to a different extent older thesis that originally were in paper format. Two of these institutions stated that they contact older students or supervisors to ask for their permission to digitize (if not already existing digitally) older thesis.

13.2.4 Who are Submitting the Theses?

At five of the seven institutions the students are the ones intended to and, the ones doing the submission of the theses to the repository, only with exceptions where extraordinary issues occur (the student being unavailable, technical difficulties etc.). At one institution it is the library staff conducting all the uploading of the theses; students are emailed and just have to allow the uploading. (This is the institution that had no plans of making uploading mandatory. In 2008 they can show to coverage of 85% of the institutions PhD theses, and very few students were said to decline the request of having their thesis in the ETD repository.) At the remaining institution the students hand in their theses on CDs and the library department upload them.

13.2.5 Experiences and Thoughts about the Content Recruitment Strategies

One of the institutions (where submission is voluntarily) thought the information sessions they held to students were an efficient and successful mean for promoting the submission of theses. It was added that it might be an inhibiting factor if the e-mail inviting to the information session came from the library instead of from the students' faculty, this from the reason that the faculty might have more "authority" than the library.

From the institutions that did operate with a mandatory policy for submission of theses, the majority could not show to any special incentives or promotional activities, nor did they seem to think such means were necessary. They simply argued that they promoted their repositories when they were new, but now that was not necessary anymore; "it has become an accepted procedure", or "uploading is required to graduation, I suppose that is enough incentive". All of these could show to a 100 % or near to it, coverage of the institution's theses.

Several of the institutions had the option of faculty or students putting an embargo on a thesis, the embargo stretching from a few months up to three years. One institution stated that they suspected many students for choosing a (long) embargo for no reasons, and that it

might not be wise to give the students such an option (this was a institutions with a mandatory policy for submission of ETD).

On institution stated that creating and managing an ETD repository is a good way to launch an institutional repository or at least to gain experience administering an institutional repository.

13.2.6 Disciplinary Differences

From the institutions that had a policy for voluntarily uploading of thesis, disciplinary differences could be spotted. Not all explicitly explained what the differences were, but from the once that did, the following was said:

- The highest share of compliance for submission is to find amongst engineering students, it is also high amongst medical and architecture students, whilst science students are average and art students the lowest.
- Good results amongst all schools at the university, except theology which is less “technology-savvy”.
- In humanities and law a printed book is the preferred format and not an e-thesis.
- Programs that have creative works are hesitant to publish works over the internet; others want to be sure they can publish in a journal before the work is available over the internet.

13.3 Findings from Investigation of the CERN-Theses Database

Below the main results from the investigation of the CERN-Theses Database will be presented. First a brief introduction of the CERN-Theses Database and the related repository CERN Document Server is given.

13.3.1 Introduction of the CERN-Theses Database and the CERN Document Server

CERN authors produce more than 2000 articles and 10 000 conference proceedings annually. A production that the CERN library seeks to collect, make accessible and preserve. In order to manage this CERN created an institutional repository already back in 1993. The repository today goes under the name the CERN Document Server²⁶ (CDS) and uses the OAI-PMH compliant software Invenio²⁷ developed at CERN. CDS currently holds over 900 000 documents, and more than 360 000 full text documents. In spite of the fact that CERN has a mandatory policy for CERN authors to submit their documents to the library (since 2001) this only happens for less than half of the known publications. The library has dealt with this

²⁶ <http://cdsweb.cern.ch>

²⁷ <http://cdsware.cern.ch/invenio>

problem, but having CDS harvest other sources. This solution has resulted in CERN having the metadata for almost 100% of their publications (but not all in full text). (Pepe and Yeomans 2006)

The CERN-Theses Database²⁸ consists as an own unit in under the CDS, and now holds about 3000 theses and dissertations produced in connection to CERN. Anyone (students, supervisor, secretariats and library staff) can submit a thesis, but after submission the library staff screens the documents, ensuring that the documents are authentic CERN related theses and dissertations.

13.3.2 CERN-Theses Database Recruitment Strategy

As already mentioned it is mandatory for a scientist at CERN to upload his work to the CDS, and the same rule applies for students; for them it is mandatory to upload their theses to the CERN-Theses Database.

The library at CERN promotes their services, including the CERN-Theses Database, at general information sessions about CERN that are held to new students coming to CERN.

13.3.3 Motives for Uploading

Some submit theses merely because it is a part of their working instructions. This applies to the secretariats at the experiments and accelerators, which aid the related scientists at making all their material available through CDS to a differing extent depending on the working habits within the respective collaboration. Several of the secretariats did not have the same (or any) routines for the uploading of theses as they have for other scientific papers, seemingly thinking that a thesis is of inferior quality to other types of scientific papers. However there was one secretariat that had defined and explicit routines for how to deal with their theses. They had lists over all the students and theses that were to be written in relation to their collaboration, and the theses due date. When the theses were supposed to be finished, the secretariat crosschecked whether it had been uploaded to the CERN-Theses Database, and if not they contacted the students and made sure it was done, the majority of times ending up by uploading the thesis themselves with the permission from the student.

The supervisors that had submitted several theses turned out to be at the Scientific Information Policy Board at CERN, and thus were very well informed about CERN's repository, their quest for Open Access and the benefits of Open Access literature. From this reason he naturally enough promoted to his students that they should upload their theses, and if they did not he uploaded them himself.

²⁸ <http://cdsweb.cern.ch/collection/CERN%20Theses?ln=en>

Some of the students were advised by their supervisor to upload their theses, whilst one student was given the same advice from the person responsible for the PhD students at CERN.

Several expressed to have come across the CERN-Theses Database when searching for theses when they were in the process of writing their own. As they found the possibility of accessing other theses very useful for themselves, they also wished to contribute and help others, by uploading their own.

A couple of the students stated that their main reason for submitting their theses was to be able to professionally cite their own thesis in later work, and they thought having it in the CERN-Theses Database was a good solution.

13.3.4 Interesting Comments

One of the students expressed the benefits of having access to other theses when writing his thesis. On the contrary he was not entirely comfortable about having his thesis available at the internet, because he felt uncertain about what it could be used for by others.

Several of the interview objects were of the opinion that “everyone” are aware of /understands the benefits of having scientific material available and accessible online, and that the real problem lies within the lack of awareness of available repositories and possibilities.

One pointed out that the CERN library is by tradition accustomed to working closely with the high energy physicist, and that they are more distant to the engineers working with technology development at the experiments and accelerators.

None of the students seem to be reluctant when asked if their theses could be submitted to the repository, without any incentives being given, besides that fact that someone does it for them.

One of the student emphasized that it was a coincidence that he heard of the CERN-Theses Database, and that it was just as much of a coincidence that he had ended up submitting his thesis to it.

13.3.5 General Findings in Relation to the CERN-Theses Database

Amongst the secretariats it was a surprisingly large unawareness of the existence of the CERN-Theses Database (but they do know of CDS, and use it frequently.) From the three secretariats that were contacted, only one knew of the CERN-Theses Database, and had routines for insuring that the theses written in connection to their experiment were preserved.

It was apparent that there exist large differences within the organization of CERN from experiment to experiment, and from discipline to discipline, when it comes to working culture and the habits for sharing of information. Even though CDS is meant to be the working repository for CERN many of the different collaborations have their own repositories which they use (frequently) for the sharing and preservation of their working documents and publications. At one of the experiments which had their own internal repository, it was said that students upload their theses to the internal repository without being told to, because the repository is a tool integrated in their work routines and uploading ones thesis is just the natural thing to do. (This collaboration also awards the best thesis written each year.)

When the interviewees were asked whether it is mandatory to submit a thesis to CDS the majority hesitated and did not know whether it was or not (in spite that it has been mandatory since 2001).

In relation to the technicalities, no one found the submission itself difficult. However, some said it was hard finding the actual submission page.

Not all the students knew that there was an own section for CERN-Theses and was not sure whether they had uploaded their thesis to CDS or the CERN-Theses Database in particular.

No one stated to see any benefits in being present in the CERN repository as opposed to another repository. No one had submitted their thesis to another repository in addition to the CERN-Repository, and none of the students knew whether their home institution had a ETD repository.

14 Discussion

The questions posed in the outside of the thesis will be discussed in this chapter. Some of the discussions are merely in the light of what one have learned from the theory review, whilst in connection to best practices for content recruitment the findings from the empirical investigations will also be discussed. In the final section of this chapter practical advices for the CERN-Theses Database, based on the empirical findings, are given.

14.1 The Purpose of Different Digital Repositories

14.1.1 Accessibility, Dissemination and Enhancing Impact

Both institutional repositories and subject repositories increases the accessibility of research papers if one compares it to them only being available to subscription based journals. The same goes for dissemination and enhancing impact of research output (by reaching a large audience). But institutional repositories are more reliant on being interoperable (which now generally means being OAI-PMH compliant) with other institutional repositories and searchable through a service provider in order for a user to find papers relevant to his discipline.

14.1.2 Long Term preservation

In theory both types of repositories serves for preserving research output, but it is not unreasonable to argue that institutional repositories might be the type most likely to ensure a long term preservation due to having a host institution which can supply additional resources if needed. (If by example file formats have to be converted due to changing standards.) A low cost subject repository run on community basis might not have such resources.

14.1.3 Work in Progress

Having a repository as a place for storing work in progress should apply mostly to an institutional repository, where research collaborations and colleagues can share documents that are relevant to each other and which they are jointly working on. To what extent institutional repositories also function as a workspace at institutions is not mentioned in the literature. However one can image it to be beneficial if well organized, due to the repository becoming an item more integrated into the daily working habits of the faculty and therefore more likely to capture their attention and research output. Such a feature will probably be spotted more seldom in a subject repository for any field of a recognized size. In arXiv which is a subject repository, there is the possibility for storing versions of a document. But one can assume that thus is mostly used for the sake of rapid dissemination and wanting to make available ones results before finishing the paper or having it published, than rather being a location for work sharing and storing of work in progress.

14.1.4 A Window to the World

An institutional repository can be an institutions window to the world, where funding bodies, potential new students and employees can get to know the institutions through its intellectual output. The same thing cannot be said to be an intention of a subject repository.

14.1.5 Management of the Institutions Intellectual Output

An institutional repository can serve as tool for monitoring and managing an institutions intellectual output. The repository can by example be used to keep track of the activity (measured in output of scientific papers) of individuals or departments.

14.1.6 A Teaching Aid

An institutional repository can facilitate the sharing of teaching aids within an institution as well as be used to manage and measure teaching activity.

14.1.7 Support and Sustain Student Endeavors

An institutional repository can keep theses that are of great help to students themselves working on similar topics or theses. It can also preserve research that is easily forgotten.

Table 2 – The Purpose of Institutional Repositories Contra Subject Repositories

	Institutional Repository	Subject Repository
Accessibility	Yes (But depending on interoperability and extended search engines)	Yes
Dissemination	Yes (But depending on interoperability and extended search engines)	Yes
Enhancing Impact of Research	Yes (But depending on interoperability and extended search engines)	Yes
Long Term Preservation	Yes	Less Certain
Window to the World	Yes	No
Storing of Work in Progress	Yes (Though uncertain how common this is)	No (Or to a very small extent)
Management of an Institutions Intellectual Output		
A Teaching Aid	Yes	No
Support and sustain student endeavors	Yes	Yes, although reliant on harvesting institutional repositories

From this it seems apparent that institutional repositories serve some institution specific purposes that do not apply to a subject repository. To some extent the institutional repository also serves the more general digital repository purposes of accessibility, dissemination and enhancing research impact. The latter points are better (or truly) served if the institutional repositories are interoperable with other institutional repositories, which together are searchable in respect to subject by a good search engine. The subject repositories serve the purpose of accessibility, dissemination and enhancing research impact within any field, a not any institutional specific purposes.

14.2 Best Practices for Content Recruitment

The literature presents many practices aimed at promoting a repository and recruiting content to it. What have not been clearly established is which of these practices that are efficient and whether some are not. It can be argued that several of the promotional activities seem to be counteracts to the reasons commonly found for poor faculty uptake and thus can be assumed effect full. Still it is apparent that many repositories are struggling to gain content, so something must still be amiss.

14.2.1 Mandatory Policies

The literature was fairly clear in its recommendation of mandatory policies for submission of research material to institutional repositories. The results from the investigation of the ETD repositories support this opinion. Amongst the repositories having mandatory policies, content recruitment strategies and promotion of the repository did not even appear to be an issue. Why mandatory policies in spite of their good reputation have not become the norm at every institution seems rather strange, and possibly indicates how commitment from high up in the hierarchy at institutions is lacking and how it is not enough to merely allocate the resources for the creation and maintenance of a repository. Still it might be argued that mandatory policies across a whole organization might not be as easy to carry out as for students where submission of an ETD can be made a criterion for graduation. Perhaps institution wide mandates could be supported by means like making submission of (at least) publications to the institutions repository a criterion for the scientist receiving his research funding.

As a contradicting example to the stated efficiency of mandatory policies, CDS and the CERN-Theses Database is an example of how mandatory polices alone cannot guarantee successful recruitment. At CERN none of the interviewees were certain whether submission is mandatory at CERN (something which it is), and the CERN library rely on harvesting other sources in order to get access to their own publications. This is an indicator of how important it is that a mandatory policy is well communicated to the whole organization it is directed at, and should be sought turned into a norm within the organization.

Whilst a mandate might be the most efficient way of recruiting content to an institutional repository, a mandate might be impossible to carry out for a subject repository which thus must rely on voluntarily submission (as long as the material does not exist in institutional repositories which can be harvested). The undersigned therefore find it rather surprising

how marketing issues like for example branding is not mentioned in the literature concerning content recruitment to these repositories, with the idea that repositories can attract submissions through a name or brand if associated with for example quality content, the best search engine, the most visited etc..

14.2.2 Content Recruitment as a Result from Usage of a Repository

ArXiv is a subject repository to be admired and being able to learn from its success would be of great value. ArXiv's success seems to be accredited to its users being pioneers when it comes to employing new technology in order to improve their scholarly communication. This was presumably being driven by their demand for rapid dissemination of results prior to their publication, and the field's long lasting traditions for sharing of pre prints. Besides bringing understanding of how traditions and field specific needs can influence the adoption of a self-archiving culture, it is difficult to extract the factors from arXiv's success which can help one improve content recruitment to another repository. However it might indicate how repositories by gaining a significant coverage of a field, are secured a continuing growth in virtue of their important place in the life of its users. As Ginsparg pointed out, ArXiv is the place where a high energy physicist can find all the material he is searching, as long as the material exists. This makes arXiv an integrated part of a physicist's working habits, and hence the natural location for him to self-archive. By generalizing this one can suggest that a repository being a working tool for a scientist is a factor which positively influences content recruitment to the repository. Thinking in the context of an institutional repository it should therefore be beneficial to the content recruitment if the repository also is the working tool of the institution for safe-keeping of work in progress as well as for work sharing.

Several of the students explained that they got to know the CERN-Theses Database and decided to archive their thesis there, because they used the database to find similar theses when they were in the process of writing their own. Many of them had searched CDS and through that saw that there was something called the CERN-Theses Database. One of the internal repositories at CERN explained how everything was self-archived in their repository because everyone within the group was used to keeping all their material there. In arXiv one can see how it has become the natural location for a physicist to self-archive as this is where he turns to find all the information relevant to his field. What these three examples illustrate is the delicate interdependence between the usage of a repository in terms of finding information or as a work tool, and between successful content recruitment. Subsequently it can be argued that promotion aimed at increasing the usage of a repository (in terms of searching or as a working tool) also will have positive effects on content recruitment. Furthermore it can indicate the importance of building a name around a repository, or at least making sure that the users sees where the information he found origins from in the instances where he uses Google or other search engines. If the user clearly sees where his information comes from, it being arXiv or the CERN-Theses Database, he might later be inclined to upload his own work to this location. As Borgmann pointed out do not everyone see how the library today is coming to them, a more explicit branding could therefore be the solution (e.g. this paper is brought to you by the CERN Document Server etc.).

14.2.3 Reasons for Poor Faculty Uptake

In the literature lack of awareness is found as one of the main reasons for poor faculty uptake of digital repositories, a perception which also seem to be shared by the interviewees. All of them seemed to think that scientists do understand the benefits of having research freely available on the internet they are just not familiar with how it should be done and which possibilities their institutions offers.

Not surprisingly did small repositories internal to a group or section exist throughout the organization of CERN, a situation which probably is not unique for CERN. One of the internal repositories that the author came across during the investigation carried theses written in connection to one of the experiments at CERN. The interviewee (responsible for publications at the respective experiment) said that to his knowledge there were no special incentives or demands given to the students about uploading their thesis to the internal repository, everyone just did so automatically because the internal repository was an integral part of their working routines, and the location where all work related documents are kept. On one hand this can be interpreted as a possible common barrier to faculty adoption of an institutional repository; that the scientists already spends energy and time on placing their research material in one location and thus are not inclined to spend more time placing it in an additional location. Possibly do they not even know of the other repository, or reflect upon that their internal repository probably not is interoperable with any other repository and thus remains an isolated island. On the other hand this example also points in the direction of factors which can improve content recruitment, namely that an institutional repository can gain content by being closely integrated with the working habits of the organization.

In any case this is an issue that has to be considered when an institution is to build a repository. What internal working routines and repositories exist? Should the users of them be sought convinced (and how) into “changing” repository or should alternatively the new repository adapt to the existing once by building interoperability across the repositories and harvesting them? The author is of the opinion that the first option is the one to prefer, and that slowly streamlining the working procedures of the institution as a whole in terms of sharing of research material can have a positive effect on innovation and learning in the organization. By ignoring these challenges the institution is more than likely to fail in their quest for creating a healthy, growing and visited repository.

14.2.4 Who Should Do the Uploading?

Harvesting of content in terms of retrospective digitization of papers as well as submission of content by library staff is mentioned in the literature as a way of populating a repository. But can it be argued that content submissions by library staff might be a way of diserving the repository in the long run? The investigation of the ETD repositories also revealed how some repositories rely on library submission. The library staff contact the students which merely have to agree to have their work uploaded something which there apparently is little reluctance against. The repository following this strategy, had no plans for making submission of electronic theses mandatory and appeared to be happy with the current 85 %

coverage of PhD-Theses. Also the secretariat at CERN which had the routine of emailing students to ask for their permission (as well as access to their work) for uploading their thesis, said that everyone agreed to do it. What might be questioned in this context is whether such a strategy is an efficient usage of resources. Could the resources the library staff requires in order to manage a submission process be more efficiently used at promoting the repository and incentives for self-archiving, or could the task merely be solved by having a (well communicated) mandate? Library staff conducting the archiving involves a continuous expenditure, whilst resources spent at promotion of a repository and/or implementing a mandate might only have to be done in an introduction face, where after submissions would be self-motored through that the scientists are being accustomed to self-archive and it becoming a norm in the organization. In other words, what might seem most efficient and successful in the short run might have better long term solutions. Another reason speaking in favor of “training” the scientists to self-archive in their institutional repository is that library staff might only be able to get hold of a fraction of what the scientists are producing. Furthermore it can be argued that as long as the scientists do not become familiar with the repository through self-archiving they might also be less inclined to use the repository, resulting in a reduction of the repository’s potential benefits for the organization.

14.2.5 Promotional Activities

Information Sessions

In terms of the promotional factors that were mentioned by the ETD repositories, only one was explicitly mentioned as perceived to be successful. This was the information sessions held to new students and graduate students. However none of the students interviewed pointed out the CERN library’s information sessions as the factor making them self-archive their thesis. This probably does not mean that such information sessions are fruitless in general, but perhaps that the ones at CERN are not very effective. To suggest why this is the situation at CERN, one can mention that it might be a problem that the information is only given to the students when they are new, and therefore they might have forgotten all about the CERN-Theses Database when the time is due for them to write and hand in their thesis. It was pointed out that invitation and information about such sessions should come from the students’ institutes and/or faculties, and not from the library which might have less influence on the students. This is an indication of the importance of repository commitment from higher up in the hierarchy, than merely the library.

The limited knowledge of which promotional activities that are successful points out the need for additional research on the matter. Maybe several of the promotional activities have no effect, and therefore is a waste of resources.

Incentives

From the investigation one learned that several ETD repositories had the practices of not making students hand in paper copies (or fewer) of their theses if handing in an ETD, thus given them an incentive to self-archiving. As many students have to pay for printing their theses, this is also an economical incentive. How efficient such an incentive is, is hard to

know, but at least one can assume that it helps create good will for self-archiving amongst the students.

14.2.6 Disciplinary Differences

The investigation of cultural differences amongst disciplines regarding uploading of content to digital repositories is not of such an extent that it can indicate any recipe for how to deal with the different disciplines. It can merely build upon what already has been suggested in the literature; that disciplinary difference in terms of scholarly communication exists and that this is a consideration important to take into account and how more promotional activities in the context of content recruitment, may need to be directed towards the less “technology savvy” disciplines.

14.3 Co-Existence between Digital Repositories and Publishers

14.3.1 Is A Co-Existence between Digital Repositories and Publishers/Journals Viable?

There seem to be no common opinion in the literature regarding whether digital repositories pose a severe threat to subscription based journals, and investigations performed onto the issue has mouthed out in very different conclusions. The investigation of the usage pattern of e-prints contra the published paper within the field of astrophysics was clear in its verdicts; digital repositories pose no threat to publishers. On the contrary they can increase the visibility of journals thanks to the links they have to the published version of the papers. The investigation of librarians’ preferences of openly accessible materials compared to material from subscription based journals painted out a very different picture; if digital repositories could fill librarians’ quality criteria regarding rapid dissemination and peer reviewed content, the librarians would happily cancel their journal subscriptions.

So which of these opposing results can one believe to best reflect the actual situation, and to be right in their predictions? In spite of being conflicting, both the scenarios seem credible, as well as the reasoning about the implications. It sounds only natural that a scientist given the choice by either entering the manuscript of a paper (or the final draught) or entering the published paper, when access to both options is only a click away, chooses the published paper (for example due to reasons like a more agreeable interface, the perceived prestige of the journal etc.). But for this to be the situation, the digital repository the scientist is using have to have links to the published versions (this is the norm in arXiv, but might be different in other repositories) and even more important; the scientist have to have access to the journal through its affiliated institution. And who manages an academic institution’s journal subscriptions? The answer is often the librarians. It seems fair that they are more than happy to cancel expensive subscriptions if they can get access to the same literature through a digital repository. When seeing it this way, the threat of digital repositories to journals suddenly seem more real in spite of what users (as readers) might prefer or do by habit. However, this is again dependent on that repositories offer access to the full range of literature (thus having to majorly improve content recruitment), as well as carry peer reviewed literature which should not be delayed more then 6 months by a possible embargo

from a publishers. All these factors are not in place today, but if they come to be one could argue that the livelihood of publishers might be severely threatened.

If being able to arrange research output in a tidy and lucid way in repositories through out the world, there is still the demand for quality insurance of the content through peer review. A function that today is fulfilled by journals. By using the terms one learned about the functions which make up scholarly communication; one can say that rewarding and certification functions are not catered for by today's digital repositories. The system for rewarding could quite easily be adapted to digital repositories, for example by using metrics like downloading of the full text version of a paper, citations etc. Whilst a well functioning alternative to the certification by journal peer review does not exist and is clearly an important function. As long as digital repositories do not vertically integrate throughout the value chain of scholarly communication by taking on peer review, the journals differentiate themselves by a much wanted offering which the repositories do not have.

If suggesting that digital repositories should start integrating into the role of peer review, in this way making journals to a large extent redundant, one would have to take into account that this would entail repository administrators entering a field they have no prior experience in; representing a rather large upheaval that would not be trouble free, and add on to the list of challenges that repositories are already struggling to manage. A voice that sounds more realistic in the long run is Ginspargs, who is pointing out that the existing form of peer review might be rather outdated for the current rapid digital way of disseminating research material, and that it should be changed in order to keep track with the technological advancements and not slow down the process of scholarly communication. However, it does not seem unfair to think that this development of peer review should happen within the journals, which are experienced in doing peer review.

The process of scholarly communication has unquestionably changed much over the previous two decades, thanks to developments on the technical side. Actors within the value chain will undoubtedly feel this change, and will have to adapt to it. As the situation is today the author will say that journals provide a core product, peer review, which no one else does, and by this defend their lively hood as long as no other parties (repositories or other instances) decides to enter this arena. One could be tempted to suggest that it would be wise of journals to start changing their business focus towards peer review and quality control, and let the function of awareness (dissemination and accessibility) to subject repositories which can be said to better serve this function than what a single journal can on its own. But the direction the existence of journals and repositories will take is hard to predict and will be shaped by how funding bodies chose to use their resources, whether the journals try to "fight" the competition from digital repositories by copyright restriction and embargos, as well as to what extent digital repositories are able to recruit content and offer extended search ability across repositories.

14.3.2 Is A Co-Existence between Different Types of Digital Repositories Viable?

Whether co-existence between both institutional and subject repositories is viable is not to the author's knowledge widely dealt with in the literature, and maybe it is more appropriate to ask whether a parallel existence of both types is a wise usage of resources. Perhaps it would be more optimal directing the resources towards one or the other type of digital repositories, and splitting the task of content recruitment between the types, in terms of catering for the stakeholders needs in the best academic and cost efficient way.

One might argue that a co-existence of different types of repositories is viable as the two types co-exist today. Further on the existence of both types can be defended by how they to some extent serve different purposes (see chapter 13.1). But this structure might be far from optimal, leading to an excess amount of resources being used to capture academic research output as well as making the landscape harder to navigate through. As repositories generally are low cost offerings without direct commercial motives they do not compete in terms of revenue. What they are competing about is the pool of resources existing in the academic sector and users in terms of readers and contributors. For users as contributors to digital repositories it is desirable to only submit a paper to one repository, not having to repeat the submission to both the institutional repository of ones affiliated institution (maybe the author even have several) as well as to the subject repository/ies relevant to his field. For the user as a reader it is desirable to be able to go to one location (one repository) and find everything relevant written to his field of research.

If Viable, What Would Be the Most Natural Structure of Co-Existence?

In the previous section it was argued for why a co-existence is viable and should continue, and how a clearer division of roles would be beneficial to all stakeholders. In terms of content recruitment this could imply that a norm should be sought developed for whether self-archiving should happen in either institutional repositories or subject repositories, letting the remaining repository type rely on content recruitment through harvesting. In this way the academic community would avoid a duplication of resources spent at content recruitment, as well as that the world of repositories might seem more lucid and organized, and subsequently have a larger appeal to the scientist.

It could be argued that the location for self-archiving should be the repository type best suited at recruiting content and ensuring long term preservation of it. On those terms the literature points in the direction of institutional repositories, as they are thought of as best suited to ensure long term preservation through their stable business models as well as to recruit content through institutional mandatory policies. For this to fully work all repositories have to be interoperable. In addition there must be developed (and applied) standards for metadata, making a complete and correct harvest feasible.

An argument speaking in favor of subject repositories being the location for scientist to self-archive is that the subject repositories might be perceived as more beneficial (in terms of reaching a larger audience) to have ones material in than an institutional repository, a trend which is has been spotted amongst the physicist. However, this author thinks that the

majority of other fields will not be inclined to self-archiving voluntarily in subject repositories, and thus need to get used to the habit of self-archiving by it being made mandatory from the scientists' affiliated institutions. It would also be important to communicate that no institutional repository will be an island, and that their content will be aggregated into the subject repositories. However, as arXiv illustrate, there might be some fields that are inclined to voluntarily self-archive in a subject repository. Amongst them, as the physicists, it might be harder to uphold an institutional mandate for self-archiving as they already feel that they are self-archiving in the most sensible location.

14.4 Digital Repositories and Open Access

14.4.1 Are Digital Repositories the Mean for Achieving Open Access?

Predicting what role digital repositories will play in the Open Access movement's quest for freely, accessible research literature is a complex task. In the previous discussion about whether a co-existence between digital repositories and journals is viable, one argued that as long as these serve different purposes and attends to different functions in the value chain of scholarly communication, as they do today or even more specialized, a co-existence is viable. However, one might come to ask whether this situation changes if all journals were Open Access and not subscription based. Would the need for digital repositories in such a situation disappear? The author does not think so. Institutional repositories would still earn their right to live through the institution specific purposes they serve (preservation, management and measurement of the institutions output), whilst subject repositories as discussed are the bodies best suited for providing accessibility through their aggregation of all material relevant to their field. Furthermore the repositories can provide more rapid dissemination of results, compared to journals which peer review results before publishing them. If assuming that digital repositories and Open Access journals are apt for co-existence, the question for the supporters of Open Access boil down to where one should focus ones effort and through which means one can expect to achieve Open Access first.

The challenges for achieving Open Access through digital repositories are mainly evolving around them being able to recruit content, as well as developing standards for metadata which makes it feasible to navigate through the literature that is preserved in the repositories. The publishing industry on the other hand, is reliant on changing their business models completely in order to become Open Access. A change than can be described as vast, and which cannot be assumed to happen over night. The numbers from the literature suggest that a change to Open Access in the short run are most likely to rely on digital repositories (as 85% of academic research material is allowed to self-archive, compared to only 5% which is available through Open Access journals). All these factors indicate that Open Access in the nearest future most realistically will be seen through digital repositories, with the journals following in the later run. If one should try to predict how the landscape will change in the publishing industry, one could suggest that the journals will start changing their business models one after the other in a slow pace. With the pace of the alteration presumably to speed up as the pressure for answering to the acts of the other Open Access journals and the repositories increases, leaving the remaining subscription based journals

with no choice but to adapt to a business model which is compatible with content being freely disseminated.

However, arguing that subject repositories provide the best availability, whilst journals are best suited for providing quality insurance, is a reason for seeking Open Access through the combination of digital repositories and Open Access journals. As suggested in the previous section, a more specific division of the roles and core functions between the parts in the value chain of scholarly communication could be developed on these terms, making room for all the different players and letting them specialize in what they can do better than others.

14.4.2 What Type of Content Do We Want in Them (Peer Reviewed or Gray)?

Harnad argues that institutions can assure the long term stability of a repository, furthermore he is of the opinion that the repositories should only contain peer reviewed material. However, by carrying merely peer reviewed material a repository would forsake much of its potential value to its institution. The information overload of research material of very differing quality which Harnad fears, can presumably be avoided by relying on technical solutions for labeling content in such a way that it can be easily navigated through. Well assigned metadata as well as labels depending of what type of certification process the respective paper has gone through (journal peer review or some kind of institutional evaluation etc.), could solve some of the quality issues, as the user then can judge to himself what kind of material he finds interesting. Such an approach would also allow the digital repositories to carry gray literature as for example teaching aids and work in progress. As Cockerill has pointed out, would repositories which only carry peer reviewed content be mere duplications of journals. Furthermore they would not serve to speed up the dissemination of results by making them available prior publication.

14.4.3 Can They Manage to Get the Content?

From the literature we learned that repositories struggle to recruit content despite that 85% of all research material is not restricted from self-archiving by any copyrights, indicating how improving content recruitment is the main challenge for achieving Open Access through digital repositories. The author is of the opinion that it is feasible as long as the focus towards the issue of content recruitment increases, and is dealt with by mandatory policies and support from the managerial positions within the organizations.

14.4.4 Ensuring Quality

Today quality is ensured by peer review which the journals can provide. Harnad worries that if these journals are made Open Access through the author pays model, the quality might decrease as journals gain revenue by additional publications. But if what differentiates the journals from digital repositories, are their certifying function as well as well possibly a brand name built upon their good reputation and associated quality, it seems fair to assume as Cockerill does, that the journals will not be tempted to decrease their level of quality due to

short term economical incentives. Digital Repositories do not currently have the capabilities of providing peer review, and as argued in the chapter discussing the co-existence of publishers and repositories, they have no such experience and are perhaps better served by leaving this task to the journals which have it as a core capability.

14.5 Recommendation for the CERN-Theses Database

The two main points the author want to emphasis, is the need for a better communication and implementation of the existing mandatory policy for self-archiving at CERN, as well as a general focus on creating awareness of the Theses Database.

14.5.1 Mandatory Policy

The investigations revealed that many are unaware that CERN has got a mandate saying that everyone should self-archive their research material to CDS (and theses to the Theses Database). Naturally enough a mandate loses its effect if it is not communicated or upheld. The author is of the opinion that communicating the mandate is not something that should merely be in the hands of the library, but an issue that should be emphasized from the top down in the organization. The thought behind it being that this would reach more people as well as the message might be more effect full when coming from a higher level. In the long term the habit of submitting ones work to the CDS or the Theses-Database should become a part of the incorporated work patterns and norms at CERN. The CERN library's practice of dealing with poor submission rates from CERN scientists, by harvesting other sources and in this way recuperating the metadata of nearly all their publications, might have become somewhat of a sleeping pillow, taking away their attention for other means that could increase the direct submission of papers to CDS and the CERN-Theses Database.

14.5.2 Awareness

As mentioned earlier in the literature and also indicated by the interviewees, one of the main problems repositories are facing is a lack of awareness. In the context of CERN the author has the impression that CERN scientists are aware of CDS, but that the awareness of the Theses Database amongst students and staff/administrative personnel is much lower. How the students came to learn about the Theses database appeared as very random. To increase the recovery rate of CERN-Theses, the library could by some simple means be able to communicate the presence of the Theses Database to all its Technical and PhD students. One thing is being present at the information sessions held to new students in order to tell them about the CERN-These database and CERNs mandatory self-archiving policy. Another is by informing and instructing the "links" at CERN that are in contact with these students. Based on this investigation such suggested links would be the H&R persons in charge of the students and the secretariats of the different experiments. By instructing these to inform students about the CERN-Theses Database and how it is mandatory to upload ones work their, the information would easily reach it destination.

14.5.3 Generally for CDS

The modest investigation in form of a few interviews at CERN, gave the author a brief insight to how diverse working habits at CERN are throughout the organization. All the practices of work sharing and work storing in repositories internal to a group or section, makes the task of convincing CERN as whole to upload their work to CDS a complex task. The CERN Library is probably aware of this, but do not seem to actively address the problem. The author suggests that a task force should be created and set to look into the problem. By interviews throughout the organization, one could map the existing myriads of repositories. Furthermore the task force should suggest technical solutions, which could facilitate either harvesting from these other repositories to CDS, or see how CDS could be adapted to fulfill the needs users of the internal repositories, and how this solutions should be implemented. In the long run the author thinks it is beneficial to aim for a transition from the usages of many internal repositories at CERN, to the practice of everyone using CDS.

Section 3 - Conclusion and Further Research

15 Conclusion

15.1.1 Purpose

As a result of the discussion comparing the purpose of institutional and subject repositories, the author concludes that they each serve their distinct purposes in addition to the general digital repository purposes related to free dissemination and preservation of research material. Institutional repositories can be the tool for monitoring, capturing and promoting an institution's academic output to the world, as well as the location for sharing and storing of work in progress and teaching materials. Subject repositories are the tool best suited for wide dissemination and information retrieval. The conclusion about how they serve different purposes is an argument speaking in favor of the livelihood of both the types of digital repositories.

15.1.2 Best-Practices for Content Recruitment

In accordance with the literature, this thesis identifies mandatory policies as efficient means for content recruitment. However, from the investigation at CERN it emerges that mandatory policies lose their value if not well communicated and upheld.

In the context of content recruitment, the thesis discusses whether repository administrators should aim to train their users to self-archive research material in the institutions' repository, or whether this task could just as well be performed by library staff, or by harvesting of other sources as done at CERN. The thesis suggests that self-archiving is to be preferred, as it may represent optimal usage of resources in the long run. Despite the fact that adjusting people's habits and working patterns to the use of an institutional repository is a challenging and a task demanding extra resources, the habits will later remain in the organization as a part of its norms and traditions, and subsequent deposition of research material would be self-motored, rather than relying on submission by administrative personnel. Furthermore, administrative personnel cannot be expected to monitor and collect more than a fraction of the actual research material produced at an institution.

The investigation at CERN has revealed how existing repositories internal to a group or section can be a barrier to comprehensive deployment of an institutional repository, and a factor that should not be ignored. On the other hand it also illustrates how a repository can get content (without any effort), by adapting to the working routines of its users by also providing a location for storage of work in progress and work sharing.

15.1.3 Co-Existence

The thesis concludes that as long as the institutional repositories, subject repositories and journals serve different functions in the value chain of scholarly communication, there will be room for all of those and a co-existence is possible. However this will depend strongly on

which direction journals, repositories and funding bodies take in the future. The thesis suggests that a clearer specialization of tasks and roles would be beneficial for all parties in a co-existence, letting everyone do what they do best. The thesis identifies institutional repositories as best suited for recruiting content through mandatory policies for self-archiving, whilst subject repositories (by aggregating subject related information from institutional repositories and offering extended search capabilities) are the bodies best suited for serving a scientist's need for finding relevant information as well as disseminating research results. Journals provide the much needed function of classification in terms of peer review, and may also offer prestige in terms of academic reputation to a scientist, and based on this analysis might do wise in focusing their business model towards this offering, rather than the dissemination and preservation of knowledge which might be better tended to by other bodies.

In terms of co-existence between institutional repositories and subject repositories the thesis suggests that it is a more optimal usage of resources if the process of self-archiving is practiced in only one of the types of repositories, making the remaining type reliant on harvesting as the mean for content aggregation. In this way, excessive resources for content recruitment will not be used, and it would give the world of repositories a more structured and organized appearance, which probably would be appealing to academics disinclined to spend time self-archiving in different locations. As institutional repositories can recruit content by exercising a mandate, and given their position as the entities best able to ensure long term preservation due to their stable business models, the thesis suggest that content should be self-archived in institutional repositories, leaving subject repositories as field specific aggregators and search providers.

15.1.4 Digital Repositories and Open Access

The thesis concludes that digital repositories have the potential to play the leading role in Open Access, and that Open Access in the near future appears more likely to happen through repositories than journals. However, this will be dependant on digital repositories being able to recruit a much larger and more complete range of content than today. Also international standards for metadata must be created in order to make the harvesting and cross searching of institutional repositories a feasible task. While repositories are facing the challenge of content recruitment, journals must also adjust their business models. This cannot be expected to happen over night, but more gradually as pressure increases to adopt a business model which is compatible with the free distribution of content. As pointed out in the previous chapters one can also argue that subject repositories may be the bodies best suited to dissemination, as well as information retrieval, and hence are able to offer extra value compared to Open Access through journals. Journals, on the other hand are currently irreplaceable due to the facilities offered for peer review.

15.1.5 Implications for Future Research

As previously mentioned, existing research on digital repositories related has focused on technical matters. As a result practical examples and guidance towards repository creation

and management have evolved around choices of software and alike. The author is of the opinion that it would be interesting to see a merge of the literature and existing knowledge concerning change management and digital repositories. This because of the fact that when one is implementing an institutional repository, one is not merely facing technical issues; one has to change the working habits of a whole institution. This challenge and advice for managing it, deserves a place on the agenda.

16 Bibliography

- Andrew 2003 Trends in Self-Posting of Research Material Online by Academic Staff. (Theo Andrew) Ariadne Issue 37, 2003. <http://www.ariadne.ac.uk/issue37/andrew/>
- Bailey Jr. 2006 What is Open Access? (Charles W. Bailey, Jr.) From the book: Open Access: Key Strategic, Technical and Economical Aspects, Chandos Publishing (Oxford) Limited, 2006. Page 13-26.
- Beckett and Inger 2007 Self-Archiving and Journal Subscriptions: Co-existence or Competition? An international Survey of Librarians' Preferences. (Chris Beckett and Simon Inger) Scholarly Information Strategies Ltd., Publishing Research Consortium, 2007. http://www.publishingresearch.net/documents/Self-archiving_report.pdf
- Bianco et Al. 2007 Towards Open Access Publishing in High Energy Physics. Report of the SCOAP³ WorkingParty, BIANCO, S., et Al., CERN, 2007. <http://cdsweb.cern.ch/record/1034368/files/open-2007-015.pdf>
- Björk 2004 Open Access to scientific publications – an analysis of the barriers to change? (Bo-Christer Björk), Information Research, Vol.9, No. 2. January 2004.
- Borgmann 2002 Challenges in Building Digital Libraries for the 21st Century. (Christine L. Borgman.) From the book: Digital Libraries: People, Knowledge, and Technology. Springer-Verlag Berlin Heidelberg, 2002, Volume 2555, Pages 1-13.
- Budapest Open Access Initiative 2002 Budapest Open Access Initiative. (Leslie Chan et Al.) New York: Open Society Institute, February 14, 2002. <http://www.soros.org/openaccess/read.shtml>.
- Chan 2004 Supporting and Enhancing Scholarship in the Digital Age: The Role of Open-Access Institutional Repositories. (Leslie Chan.) Canadian Journal of Communication, Vol 29, No 3, 2004.

- Cockerill 2006 Business models in Open Access publishing. (Matthew Cockerill.) From the book: Open Access: Key Strategic, Technical and Economical Aspects, Chandos Publishing (Oxford) Limited, 2006. Pages 111–120
- Copeland et Al. 2005 Electronic Theses: the turning point. (Susan Copeland, Andrew Penman and Richard Milne) Electronic library and information systems, Vol. 39, No. 3, 2005, pages 185-197.
- Crow 2002 The Case for Institutional Repositories: A SPARC Position Paper. (Ryam Crow, A SPARC Consultant) ARL Bimonthly Report 223, 2002. http://works.bepress.com/ir_research/7
- Davis and Connolly 2007 Institutional Repositories: Evaluating the reasons for non-use of Cornell University's installation of D-Space (Philip M. Davis and Matthew J.L. Connolly, 2007) D-Lib Volume 13, Number 3/4, 2207.
- Edwards 2001 Global Knowledge: A Challenge for Librarians. (Christopher Edwards.) IFLA Journal, 2001; 21; 65. Sage Publications.
- Foster and Gibbons 2005 Understanding Faculty to Improve Content Recruitment for Institutional Repositories. (Nancy Fried Foster and Susan Gibbons) D-Lib Magazine January 2005. Volume 11 Number 1.
- Genoni 2004 Content in Institutional Repositories: A Collection Management Issue. (Paul Genoni) Library Management, 2004, Vol. 25, Issues 5/6, page 300-306.
- Gentil-Beccot et Al. 2008 Information Resources in High-Energy Physics. Surveying the Present Landscape and Charting the Future Course. (Anne Gentil-Beccot, Salvatore Mele, Annette Holtkamp, Heath B. O'Connell, Travis C. Brooks, 2008) [arXiv:0804.2701v2](https://arxiv.org/abs/0804.2701v2)

Ginsparg 2004	Scholarly Information Architecture, 1985-2015. (Paul Ginsparg) Data Science Journal, Volume 3, 4 February, 2004.
Guédon 2004	The “Green” and “Gold” Roads to Open Access: The case for Mixing and Matching. (Jean-Claude Guédon.) Serials Review, Volume 30, Number 4, 2004. Elsevier Inc.
Guédon 2006	Open access and a symption and a promise. (Jean-Claude Guédon) From the book Open Access: Key Strategic, Technical and Economical Aspects, p. 27-38.
Harnad 2008	Approaches to Open Access in Scientific Publishing (J. Harnad, 2008.) arXiv:0811.2603v4
Harnad and Brody 1991	Stevan Harnad and Tim Brody 1991, http://www.authorstream.com/Presentation/aSGuest440-92338-research-impact-cycle-self-archiving-harnad-news-reports-ppt-powerpoint/ , slide 45: “Don’t conflate the different forms of institutional archiving”.
Henneken et Al. 2006	E-prints and Journal Articles in Astronomy: a Productive Co-existence. (Edwin A. Henneken, Michael J. Kurtz, Simeon Warner, Paul Ginsparg, Guenther Eichhorn, Alberto Accomazzi, Carolyn S. Grant, Donna Thompson, Elizabeth Bohlen and Stephen S. Murray, 2006) arXiv:cs/0609126v1
Jones and Andrew 2005	Open Access, open source and e-theses: the development of the Edinburg Research Archive. (Richard Jones and Theo Andrew) Electronic library and information systems, volume 39, number 3, 2005.
Jones, Andrew and MacColl 2006	The Institutional Repository. (Richard Jones, Theo Andrew, John MacColl.) Chandos Publishing (Oxford) Limited, 2006.
King and Alvarado-Albertorio 2008	Pricing and other means of charging for scholarly journals: a literature review and commentary. (Donald W. King and Frances M. Alvarado-

	Albertorio.) Learned Publishing, Vol. 21, No. 4, page 248-272, October 2008.
Kling and McKim 2000	Not Just a Matter of Time: Field Differences and the Shaping of Electronic Media in Supporting Scientific Communication. (Rob King and Geoffrey McKim, 2000) http://rkcsi.indiana.edu/archive/CSI/WP/wp99_02B.html
Kreitz et Al. 1996	The Virtual Library in Action: Collaborative International Control of High-Energy Physics Pre-Prints. (P. A. Kreitz, L. Addis, H. Galic and T. Johnson) SLAC-PUB-7110, February 1996.
Kreitz and Brooks 2003	Subject Access through Community Partnership: A Case Study. (Patricia A. Kreitz and Travis C. Brooks, 2003.) arXiv:physics/0309027v1
Lynch 2003	Institutional Repositories: Essential Infrastructure for Scholarship in the Digital Age. (Clifford A. Lynch) ARL, no. 226, February 2003, pages 1-7. http://www.arl.org/resources/pubs/br/br226/br226.ir.shtml .
Lynch and Lippincott 2005	Institutional Repository Deployment in the United States as of Early 2005. (Clifford A. Lynch and Joan K. Lippincott.) D-Lib Magazine, September 2005, Volume 11, Number 9.
Mark and Shearer 2006	Institutional Repositories: A Review of Content Recruitment Strategies. (Timothy Mark and Kathleen Shearer, 2006) www.ifla.org/IV/ifla72/papers/155-Mark_Shearer-en.pdf
McMillan 1999	The Evolving Genre of Electronic Theses and Dissertations. (Gail McMillan) The 32nd Hawaii International Conference of Systems Science, 1999.
Pepe and Yeomans 2006	Protocols for Scholarly Communication (Alberto Pepe and Joanne Yeomans, 2006) arXiv:cs/0611005v1
Pinfield 2004	A mandate to self-archive? The role of Open Access institutional repositories. (Stephen Pinfield 2004) Based on a paper presented at the UKSG seminar

'Scientific Publications: Free for all?', The Geological Society, London, Tuesday 23 November 2004

- Roosendaal and Geurts 1997 Forces and functions in scientific communication: an analysis of their interplay. (Hans E. Roosendaal and Peter A. Th. M. Guerts.) Cooperative Research Information Systems in Physics, August 31-September 1997, Oldenburg, Germany. www.physik.uni-oldenburg.de/conferences/crisp97/roosendaal.html
- Sale 2006 The Impact of Mandatory Policies on ETD Acquisition. (Arthur Sale) D-Lib Magazine, April 2006, Volume 12, Number 4.
- Stueart 1989 Educating Information Professionals for the 21st Centurt. (Robert D. Stueart.) <http://web.simmons.edu/~chen/nit/NIT'89/89-389-stuart.html>
- Swan and Brown 2005 Open Access Self-Archiving: An author study. (Alma Swan and Sheridan Brown, m2005) <http://eprints.ecs.soton.ac.uk/10999/>
- Swan 2006 Overview of scholarly communication. (Alma Swan) From the book: Open Access: Key Strategic, Technical and Economical Aspects, Chandos Publishing (Oxford) Limited, 2006. Pages 3 -12.
- Swan 2008 The Business of digital repositories. (Alma Swan 2008) From the report DRIVER's Guide to European Repositories, Amsterdam University Press, 2008, pages 15-47.
- Van de Sompel 2004 Rethinking Scholarly Communication – Building the System that Scholars Deserve. (Herbert Van de Sompel, Sandy Payette, John Erickson, Carl Lagoze and Simeon Warner.) D-Lib Magazine September 2004. Volume 10, number 9. <http://www.dlib.org/dlib/september04/vandesompel/09vandesompel.html>
- Van Westrienen and Lynch 2005 Academic Institutional Repositories; Deployment in 13 Nations as of Mid 2005. (Gerard van Westrienen

and Clifford A. Lynch.) D-Lib Magazine, September 2005, Volume 11, Number 9.

Warner 2003

Eprints and the Open Archives Initiative. (Simeon Warner, 2003.) [arXiv:cs/0307008v1](http://arxiv.org/abs/cs/0307008v1) Library Hi Tech, Volume 21, Number 2, 151-158 (2003)

Internet Resources

<http://arxiv.org/>
<http://www.library.uiuc.edu/scholcomm/glossary.html#d>
http://www.open_doar.org/
<http://www.diva-portal.org/smash/aboutdiva.jsf>
<http://eprints.whiterose.ac.uk/docs/information.html>
<http://arrow.edu.au/about/>
<http://eprints.whiterose.ac.uk/docs/information.html>
<http://www.ndltd.org/>
<http://www.sherpa.ac.uk/projects/>
<http://www.driver-repository.eu/>
<http://www.ethos.ac.uk/>
<http://www.openarchives.org/OAI/OAI-organization.php>
<http://www.oaforum.org/tutorial/index.php>
<http://www.openarchives.org/Register/BrowseSites>
<http://www.soros.org/openaccess/view.cfm>
<http://oa.mpg.de/openaccess-berlin/berlindeclaration.html>
<http://scuri.ac.uk/WG/OATS/institutionalsupport.htm>
<http://www.soros.org/openaccess/read.shtml>
<http://www.earlham.edu/~peters/fos/bethesda.htm>
<http://www.wellcome.ac.uk/About-us/Policy/Policy-andpositionstatements/WTD002766.htm>
<http://www.oecd.org/>
<http://www.eprints.org/openaccess/policy/signup/>
<http://www.itu.int/wsis/index.html>
[http://open-access.web.cern.ch/Open-Access/pp.html\(http://scoap3.org/](http://open-access.web.cern.ch/Open-Access/pp.html(http://scoap3.org/)
<http://scoap3.prg/>
http://arxiv.org/show_monthly_submissions
<http://public.web.cern.ch/public/en/About/WebStory-en.html>
<http://www.slac.stanford.edu/spires/about/>
<http://scholar.google.ch/>
<http://www.sciencedirect.com/>
<http://www.springerlink.com/home/main.mpx>
http://www.open_doar.org/
<http://cdsweb.cern.ch>
<http://cdsweb.cern.ch/invenio>
<http://cdsweb.cern.ch/collection/CERN%20Theses?ln=en>

17 Appendix

17.1 Glossary

ETD: An electronic thesis and dissertation. It can refer to any thesis in electronic format, but it is especially intended for born-digital theses.

E Print: An electronic version of an academic's research output. The term generally includes pre- and post prints of journal articles, conference, and working papers.

Gray literature: A term used to define scholarly material which is not formally published and is either difficult to find or generally ephemeral. Examples include electronic thesis and or technical reports.

Metadata: Information used to describe an object; in this context usually in a pre-specified language; literally; data about data. For librarians, metadata creation is a generic term for the activity they traditionally describe as 'cataloguing'.

Pre-Prints: A copy of a journal article which has not been peer reviewed. Sometimes publishers will permit this copy and not the post-print (copy of journal article after peer-review and possibly publication) to be the one deposited in an institutional repository.

(These definitions are recounted from Jones, Andrew and MacColl 2006)

17.2 The ETD Repositories in the Investigation

Table 3 – ETD Repositories in the Investigation

Name of the Repository	Web Address	Host Institution
eScholarship@BC Repository	http://escholarship.bc.edu/dissertations/	Boston College
Hong Kong University Theses Online	http://sunzi1.lib.hku.hk/hkuto/index.jsp	Hong Kong University
SUNeTD	http://www.lib.sun.ac.za/Library/eng/finding/theses.html	Stellenbosch University
E-Thesis	http://ethesis.helsinki.fi/en/	Helsinki University
Electronic Theses and Dissertations at NCSU Libraries	http://www.lib.ncsu.edu/ETD-db/	North Carolina State University Libraries
MSpace at University of Manitoba	http://mspace.lib.umanitoba.ca/handle/123456789/5	Manitoba University
Electronic Theses and Dissertations at Birmingham University	http://etd.byu.edu/	Birmingham Young University

17.3 The Questionnaire sent by e-mail to the ETD Repositories

Questions Concerning Electronic Theses and Dissertations (ETD) Repositories

The following questions are made in order for the undersigned to get a better insight to the challenges and success factors related to creating and managing a vivid ETD repository. For all the questions it is of great interest that you elaborate on the topic and share what you have experienced, even though it might be on the side of the core of the question and suggested answers. Anything is of interest and the questions/ suggested answers are not meant to be limiting your answers, it is rather meant as a kick start.

Basic Facts

1. What is the repository's size?
 - a. Number of ETD's?
 - b. Percent age coverage of EDT's?
2. What is the repository's year of birth?
3. Is the ETD repository cooperating with any other ETD repositories or participating in an ETD repository initiative (for example a national ETD preservation initiative)?

- a. If so, please elaborate on the extent of this relationship (Advantages, experiences etc.)

Institutional Policies

- 4. What is the institutions policy in terms of the uploading of ETD's?
 - a. Mandatory
 - i. If not, plans for making it mandatory in the near future?
 - b. Voluntarily
 - c. Automatically done
- 5. What are the reasons for choosing such a policy?

Content Recruitment Strategies

We wish to learn what your strategies for creating awareness of your repository and for recruiting EDT's are/was, and how successful you perceive them to be/have been. It is also interesting to hear under what assumptions these strategies have been made. Any elaborations on challenges/successes/experiences you have had in this relation are greatly appreciated.

- 6. Who is the part intended to upload the ETD's?
 - a. Students
 - b. Library or Institutional staff
 - c. Other?
- 7. Has there been any type of general promotion and awareness creating of the repository? Like for example:
 - a. Passing out brochures
 - b. Conducting presentations
 - c. Advertising or publishing articles in the library/institutional newspaper
 - d. Formally launching of the repository at its birth?
 - e. Communication of benefits related to uploading your work in the repository?
 - i. Which benefits are communicated?
 - ii. How are they communicated?
- 8. Are there given any special incentives in order to ensure the uploading of ETD's?

Results

- 9. Who performs the majority of the uploading of the ETD's?
 - a. Students
 - b. Library or Institutional staff
 - c. Other?
- 10. Is the uploading of ETD's living up to expectations?

11. Are there any major differences in rate of successful uptake between members of different disciplines?

Content Harvesting

12. Have there been any activities, besides strategies to ensure author based content uploading, to fill the repository? Like for example;
 - a. E-mailing authors directly, requesting them either for the permission to upload their ETD's or encouraging them to do so themselves?
 - b. Cooperation with research groups, where they upload their students EDT's or encourage them to do so?
 - c. Digitization and uploading of past theses and dissertations.

Other

13. Any other opinions/advices/experiences you would like to share on the topic of ETD repositories?

17.4 Templates for the interviews with users of the CERN-Theses Database

17.4.1 Questions for Student

1. How did you hear about the CERN-Theses Database?
2. Is/was anyone giving you incentives or instructing you to submit your thesis to the CERN-Theses Database?
 - No
 - Yes, if so who:
 - My supervisor
 - My experiment/ research group
 - The library staff
 - Other students
 - Other, if so who?
 - What incentives?
3. What is your affiliated institution (the one you wrote your theses in connection to?)
4. Does your affiliated institution have an Electronic Theses and Dissertation (ETD) repository (database)?
 - Yes, if so
 - o Did you upload your theses to this repository?
 - Yes, is so why?
 - No, if so, why not?
 - No
 - I don't know
5. "When writing you theses in relation to CERN, it is mandatory to submit it to the CERN-Theses Database".

- True
 - False
6. What do you perceive as benefits from uploading your thesis to a repository (database)?
 - Safe preservation
 - My work becomes accessible for others
 - My work might impact others
 - My work will be more easily recognized
 - None
 - Other. Is so, please elaborate.
 7. Do you perceive there to be any particular benefits involved in having your thesis in the CERN-Theses database, compared to another repository?
 - No
 - Yes. If so, please elaborate.
 8. Do you know of any other repositories (“databases”)?
 - No
 - Yes. If so, which?
 9. Have you uploaded your thesis (or other work) to any other repositories/databases?
 - No
 - Yes. If so, which?
 10. Did you find the CERN-Theses submission function intuitive and easy to use?
 - Yes
 - No
 11. What parts of the submission application do you think should be improved?
 - The listed alternatives of experiments and accelerators/facility
 - What other alternatives?
 12. Are you satisfied with the CERN-Theses database?
 - Yes
 - No
 13. Are there any changes of the CERN-Theses database that you would like to see?
 - No
 - A more intuitive and easy submission process
 - The appearance of the website
 - The accuracy of the content
 - Other? Please, elaborate.
 14. How often do you consult the CERN-Theses database?
 - More often than once a week
 - Once a week
 - A couple of times a month
 - More seldom than a couple of times a month
 - Never
 15. Any comments on the topic of repositories and the CERN-Theses database or how you learned about CERN-Theses Database or why you chose to upload your work there, is greatly appreciated.

17.4.2 Questions for Supervisor

1. How did you hear about the CERN-Theses Database?
2. Why have you uploaded theses to the CERN-Theses Database? Is/was anyone giving you incentives or instructing you to submit your students' theses to the CERN-Theses Database?
3. What is your affiliated institution?
4. Do you know of any other repositories ("databases")?
5. Have you uploaded theses to any other repositories/databases?
6. Did you find the CERN-Theses submission function intuitive and easy to use? Any comments on the topic of repositories and the CERN-Theses database or how you learned about CERN-Theses Database or why you chose to upload the theses there, is greatly appreciated.

17.4.3 Questions for Secretariat

1. What are your routines for preservation of theses written in connection to your experiments?
2. You wrote that you contact students and invite them to upload their theses to CDS. Does everyone approve of this, and answers yes to your request?
3. When inviting them to upload their theses to CDS, are you putting forward any arguments for why they should do so?
4. Is it you who perform the uploading of the theses the majority of the time or is it the student himself?
5. Is it a mandatory policy that material produced in relation to your experiment is uploaded to CDS?
6. Is submission of theses (and other research material) a part of your work instructions?
7. Are you given any other incentives by someone in order to ensure that all submitting of material is done?
8. Are you uploading the theses to any other repository or archives as well?

17.5 Original Scope of Thesis (Oppgaveteksten)

Norwegian University of
Science and Technology
NTNU

Faculty of Social Science and
Technology Management



Department of Industrial Economics and Technology Management

MASTER THESIS

For

Stud.techn. Reidun Anette Hagen

Field of study	Industrial Management Bedriftsadministrasjon
Start date	11.02.2009
Title	Digital Repositories in Research Based Institutions and Communities. Digitale oppbevaringssted i forskningsrelaterte institusjoner og fagområder.
Purpose	To investigate the appropriateness of repositories in research based institutions versus repositories for academic fields. The thesis will be focused towards institutions in, and the community of, the field of high energy physics. The paper will aim at developing an understanding of how to successfully ensure user adoption and content recruitment of a research related repository.

Main contents:

The emergence of repositories within research based institutions and fields of research, seen in the context of the purpose of libraries, publishers and the Open Access movement.
What are the success factors for achieving a healthy repository, that is constantly growing and being used?

Tim Torvatn
Department Management

Roger Sørheim
Supervisor

17.6 Self-Declaration Form (Egenerklæringsskjema)

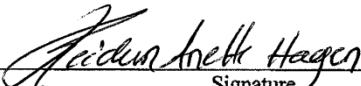
DECLARATION

Stud.techn. REIDUN ANETTE HAGEN
Institute of Industrial Economy and Technology Management

I hereby declare that I have written the above mentioned
thesis without any kind of illegal assistance

Geneva, Switzerland
Place

25.06.09
Date


Signature

In accordance with Examination regulations § 20, this thesis, together with its figures etc., remains the property of the Norwegian University of Science and Technology (NTNU). Its contents, or results of them, may not be used for other purpose without the consent of the interested parties.