Electronic Resources and Institutional Repositories in Informal Scholarly Communication and Publishing

Isabel *Galina Russell*

A thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosophy

Department of Information Studies

University College London

June 2009

Declaration of originality

I, Isabel Galina Russell confirm that the work presented in this thesis is my own work and that, to the best of my knowledge and belief, it is compromised of original research and ideas, except as acknowledged in the text.

Signed: Isabel Galina Russell

PhD Candidate

Abstract

The aim of institutional repositories is to aid the management and dissemination of the increasingly copious amount of scholarly electronic resources produced by academics. To date most research has focused on the impact for formal scholarly publishing. The purpose of this exploratory study is to discover the impact of IRs on the visibility and use of digital resources with particular focus on resources outside the formal publishing framework. An online survey and interviews with repository managers were conducted. A link analysis study was undertaken to determine what types of web resources were linking to items within repositories.

The findings show that a wide range of non-formal e-resources are accepted and repository managers' attitudes are positive towards their importance. In practice the range of resources is limited and mainly text based. The development of typologies for non-formal resources is done in an *ad hoc* manner. Workflow processes for content acquisition in repositories vary considerably and are quite complex in particular for non-formal e-resources. The findings show a lack of cohesive discourse between repository objectives and collection policies and actual work flow processes. Repository managers consider usage data important and its most popular uses are for advocacy and securing funding. Interpretation of usage data focuses on formal resources but evidence suggests that non-formal resources play an important part in repository visibility. Blogs, academic pages and discussion forums are important web sources that link to items within repositories.

The study demonstrates that institutional repositories are not particularly successful at handling resources outside the framework of formal publishing. The system caters largely towards e-prints, in particular postprints. A fundamental challenge, if scholarly communication is to move towards new forms of communication and publishing enabled by digital technologies, is to find ways to effectively name, manage and integrate non-formal electronic resources into the institutional repository.

Acknowledgements

I would like to thank my supervisor Dave Nicholas for his support during the undertaking of this research. He provided me with helpful comments for my work as well as suggestions and contacts to talk to. In particular he granted me many teaching and research opportunities during my time at the Department of Information Studies for which I am grateful.

A very special thank you to my second supervisor Claire Warwick for the time and interest she invested in my research. I am thankful not only for her shrewd and insightful remarks but also for reminding me to believe in myself when things got too overwhelming. She and Melissa Terras also gave me the opportunity to learn important research and networking skills during my time on the LAIRAH project which proved indispensable when carrying out my own work. I also appreciate both of them for trusting me with their classes during leaves of absence. Teaching was a wonderful experience for me.

I would especially like to thank Dr. Hamid Jamali who showed me the ropes from my first day at UCL and who was always willing to help me out with my hundreds of questions. I am extremely grateful for all your help. Also to my fellow PhD students/office mates Nikoletta Pappa, Jon Rimmer and Mehrnoush Mozaffarian not only for our talks about research but also for sharing a laugh and making my days at the office much happier. I would also like to thank everyone from the Department of Information Studies who made these years a pleasant experience. As a foreign student I must say I never felt lonely. A special thank you to David Clark and other members of the CIBER group.

I was lucky enough to be part of a larger community and in particular I would like to thank the Cybermetrics Group at the University of Wolverhampton for providing a home away from home, Isidro Aguillo for his enthusiasm for Cibermetría and Jutta Haider from City University for the helpful chats and overall encouragement.

A big thank you to all the repository managers who agreed to be interviewed and who took time out of their busy schedules to talk to me and answer my emails.

I am deeply indebted to Alejandro Pisanty and Juan Voutssas whose continuing support allowed me to do this PhD in the first place. I am grateful for this opportunity to study in the UK and I hope to be worthy of the trust that they, and the UNAM placed in me. I hope that upon my return to Mexico I will be able to put to good use the knowledge and experiences I have gained in order to contribute to the progress of my country. I would also like to thank the CONACYT for funding my studies.

Finally I would like to thank my family and close friends for continually supporting and believing in me and for not asking too frequently how my thesis was going and when was I going to finish. To my friends in Mexico and other countries who encouraged and supported me long distance (and even came to visit sometimes!) To my friends in London: Alejandra del Río, Alfonso Vega, Ennio Michelis, Ernesto Priego, Isabel Villaseñor, Natalia Best, Marina Milidoni, Bridget Savage, Harry Warner and Vicky McDougall for keeping me sane and happy. To UR & AL (Robert and Elizabeth Russell) for their continuing support and interest, to the Savage family for feeding me lovely homemade Sunday lunches and Kaarina Meyer for providing a London home. A particular thank you to my mother Jane Russell who carefully proof read several chapters and whose comments were particularly useful and to my father Carlos Galina whose advice on what the English are like helped me through some rougher times.

I would like to dedicate this thesis to my brother Samuel Galina who always phoned and was a constant source of encouragement and laughter when I really needed it and to Emiliano Zolla my biggest critic, friend, supporter and fan and with whom this whole adventure began.

Table of contents

List of Figures

List of Tables

Chapter 1 - INTRODUCTION

Electronic publishing and new types of digital resources have had an impact on the traditional scholarly communication system and are changing our notions of what scholarly publishing means by modifying the way researchers produce, communicate and access information. The advent of electronic publishing has been heralded as breakthrough technology that can revolutionize the way that the academic community communicates and publishes research. Some print based formats such as journal articles have migrated quite easily to the digital world. However, these types of formal publications are only a part of the universe of online digital academic resources. The online world has created the possibility for a broad range of diverse digital academic resources to be made available through the Internet. Universities are finding new ways to capture, manage and disseminate these scholarly electronic resources and institutional repositories have been proposed as a tool to aid academics to manage and distribute their digital materials. This universe is still fairly unexplored and we are not certain on exactly what diverse types of electronic resources are available nor what their different characteristics are. Additionally it is still not clear if these resources are being used and what for.

This thesis is a study on the impact of institutional repositories on the visibility and use of electronic resources with particular focus on digital types that are outside the framework of formal electronic publishing. The research is a mixed-methods study that uses both qualitative and quantitative approaches. An online survey for repository managers was conducted followed by interviews with seven case study repositories. In order to shed further light on the use of digital resources a link analysis of case study repositories was carried out to determine what types of web resources were linking to items within repositories.

The purpose of this introductory chapter is to outline the statement of the problem and the motivations for this study by indicating its importance and relevance for the field of Information

Studies. The aims and objectives of this research are explained and the research questions defined. An outline of the thesis is also presented.

Statement of the problem

An important outcome of the digital environment has been the creation and publication of digital scholarly materials on the Internet by members of the academic community that in turn has led to a growth in the amount and variety of electronic resources available online (Greenstein and Trant 1996). Potentially the web offers authors the possibility of a low cost threshold entry point to a global academic (and non-academic) audience. However, simply putting things online does not necessarily mean that these resources will have the necessary visibility in order to be found by potential readers, and many valuable resources are for all practical purposes invisible (Dunning 2006).

Universities have since developed different ways of managing, disseminating, detecting and providing access to this increasingly copious amount of material. In recent years, institutional repositories have been proposed (Harnard 2001; Crow 2002; Hubbard 2003; Lynch 2003) as a tool for digital resource management and dissemination. Although there are different types of scholarly repositories (Heerey and Anderson 2005), they all share the underlying motivation of improving access and visibility of these digital resources in order to aid their discovery and use by other scholars (Chan 2004; Heery and Anderson 2005; Kircz 2005; CENLLFEP Committee 2006). Some repositories also address the issue of digital resource preservation. It is therefore increasingly relevant to understand the design and use of these repositories and to find appropriate methodologies to determine and evaluate the impact they are having on the use of electronic resources.

In particular it is important to study the resources found within repositories that are outside the framework of traditional scholarly publishing, as there is a limited understanding of their function and impact within the scholarly communication system. In the formal publishing arena there has been extensive research regarding the motivation and mechanisms that lead academics

to communicate and publish the results from their work. In the case of electronic resources, which are outside the formal framework, there has been little research into the way that researchers are producing, communicating and accessing this information (Houghton, Steele et al. 2004). The digital environment not only allows researchers to disseminate traditional informal resources such as reports and working papers but it also opens the possibility for the development of new kinds of scholarly materials. There has been less research into the possible impact of novel types of digital genres on scholarly communication and these could have implications for the way scholarly communication and publishing work. As more research is published online in digital forms it would be important to relate this to sociological aspects, including reward systems and prestige, to further understand researcher's communicative and publishing behavior (Kling, Rosenbaum et al. 2005).

Additionally the traditional divisions of formal and informal scholarly communication and publishing are being altered and blurred with the introduction of networked computers and in particular the web (Meadows 1998; Kling and McKim 2000; Ramalho and Castro Neto 2002). Many discussions surrounding electronic publishing tend to tacitly assume that these channels are unaltered, by taking what is normal in the print world and applying the same criteria to the electronic world (Kircz 2002). However, it is quite clear that there are implications and these should be accounted for in order to better understand the current state of electronic publishing and communication, as well as their future.

Motivations for this study

Research on digital publishing has, in the first stages, focused very much on access and distribution to formal publications. But what is happening with resources that are outside this formal framework now that access and distribution channels are changing? Are researchers producing new types of materials? Does this make them communicate and research in new ways? There are three main key issues motivating this study:

a) A need for further insight into the institutional repository landscape and to evaluate the effectiveness of repositories to manage and disseminate academic research output, in

particular resources outside the framework of formal electronic publishing. What types of resources are actually stored within repositories? Initial studies seem to indicate that there is a wide range of materials (Ware 2004a; Lynch and Lippincott 2005; Westrienen van and Lynch 2005; McDowell 2007; Rieh, Markey et al. 2007; Barkier 2008) but how are these being managed? How effectively is this being done? What impact will repository management have on the types of resources made available and will this imply changes in the future for scholarly communication?

b) Address the issue of building a more sophisticated digital genre vocabulary. Usually informal academic electronic sources tend to be referred to as "everything else that is not articles and books". There are limitations when we attempt to study such an extensive spectrum of material underneath broad terminology, particularly if we wish to focus in more depth on their impact. If these resources could be new vehicles of scholarly communication it is important that we expand our typology vocabulary. In order to do this an important initial step is to look in more detail at the different types of resources available in institutional repositories. By not placing heterogeneous digital sources all together underneath one category, we can differentiate between them and better understand their particular characteristics. In the same manner that it has been argued that there are subject differences between different types of publications, it is highly likely that digital genres will also have different uses and values for different subjects. A first step is to begin to have a closer look at the resources currently available.

c) The need to understand how these new digital resources can be used to work towards Ciberinfrastructure (or eScience) and more sophisticated uses of digital medium for research and teaching. These initiatives (Hey and Trefethen 2005) are examining not only the technical aspects of information and technological systems, but also the way that information is produced, manipulated, exchanged, disseminated and used by

researchers. For example, the Cyberinfrastructure initiative in the USA, hopes to build "more ubiquitous, comprehensive digital environments that become interactive and functionally complete for research communities in terms of people, data, information, tools, and instruments and that operate at unprecedented levels of computational, storage and data transfer capacity" (Atkins, Droegemeier et al. 2003:12). New and more powerful technologies require information and research results in a more flexible format that will allow for the reuse of data for different purposes. These types of innovations will most likely make use of diverse electronic resources and will result in the further blurring of the boundaries between formal electronic publishing, grey literature and other types of electronic resources. It seems almost inevitable that electronic publishing will continue to change and to challenge our notions of publication and the scholarly communication process in general. It is therefore important to provide further insight into the different types of electronic resources that are being produced and used, in order to better understand what implications they may have for the near future of scholarly communication and publishing.

Aims and objectives

The primary aim of this thesis is to investigate the role of institutional repositories in the visibility and use of electronic resources with particular focus on digital resources outside the framework of formal electronic publishing. These will be referred to as non-formal e-resources. As the networked research environment allows academics to produce, communicate and access information in different ways this research aims to understand the role and characteristics of new digital resources within institutional repositories. The work will complement existing research on repositories and formal publishing and suggest the implications for scholarly communication and publishing in general.

The objectives are:

- To critically examine repository approaches towards collecting and disseminating non-formal electronic resources. This study will attempt to establish the criterion involved and in particular the attitudes both explicit and implicit towards the value, function and importance of non-formal resources. This will help towards understanding the possible role of institutional repositories for managing and disseminating new types of digital resources required for future more e-science type research.
- To explore new methodological approaches for measuring and evaluating use of electronic resources within institutional repositories. In particular link analysis as a methodological approach to understanding resource visibility will be evaluated, and the potential and limitations explored. Additionally the use of usage data by repository administrators will be examined, focusing on their attitudes towards its perceived usefulness and how the data is interpreted.
- To investigate in more depth the range and distribution of non-formal electronic resources within institutional repositories. The study also assesses the limitations of current digital typologies for understanding their role.
- To discuss the implications for informal and formal publishing and communication, in particular with regard to notions of scholarly publishing in order to predict if these are going to change our notions of communicating and publishing.

Main research questions

In order to achieve these aims the following research questions will be answered.

- How are non-formal electronic resources managed in institutional repositories?
- What are the attitudes towards non-formal electronic resources amongst repository managers?
- What are the different types of non-formal resources within institutional repositories?
- What is the distribution of non-formal resources within institutional repositories?
- To what extent are electronic resources in repositories being used?
- What methodology can be used to evaluate the use of electronic resources in repositories?
- What are the implications of non-formal electronic resources in institutional repositories for scholarly communication and publishing?
- How can the boundaries between formal and informal publishing be defined in a new repository-based environment?

Scope

This study focuses on non-formal electronic resources contained within institutional repositories. This section defines the use of the term non-formal electronic resources and institutional repositories. Additionally both terms are addressed in more detail in the literature review.

Non-formal electronic resources

For the purpose of this study electronic resources that are outside the framework of formal electronic publishing can be divided into two groups. The first group is made up of documents that have an obvious print counterpart. Examples of these types of materials are theses, conference proceedings, reports, working and white papers. These types of materials are usually considered to be on the borderline between formal and non-formal types of communication. Their acceptance as a formal publishing channel varies between disciplines. For example, computer scientists tend to accept conference proceedings as a publication venue whilst biologists may see them as a more informal form of communication (Kling and McKim 1999).

In the traditional print world these types of documents were generally regarded as 'grey literature'. These are publications where the producing body's main activity is not publishing. For example, working papers produced by an academic department or an organization's newsletters. These types of publications are usually difficult to locate and access, as they do not go through the main publishing channels. There has been little or limited research on the impact, use and dissemination of these types of materials in the electronic world, despite a general agreement that these materials are important for academic research (Banks 2005).

The second group refers to the new types of resources that do not have an equivalent print format and that have become increasingly important in communication and publishing. Examples of these types of materials are blogs, podcasts, image databases, tagged databases of resources and a wide variety of web pages that use technology to present information in new and varied ways. In strict and traditional sense these types of materials would not be regarded as formal scholarly publications, although it is quite clear that they can be used for scholarly communication and as a means of disseminating results. It is therefore important to evaluate them and to conduct further research into the creation and use of these electronic resources, both for the previously mentioned electronic grey literature as well as these new types of electronic resources. It is important to note however, that boundaries in electronic publishing are not clear-cut and these distinctions may not always be possible. In addition, it is highly likely that there will be disciplinary differences.

This wide range of electronic resources makes it particularly difficult to determine their use as we are still not very sure what types exist within repositories and their relative distribution and importance within the repository. An important area of work for this thesis was to develop a broad classification of the different type of electronic resources contemplated within the repository in order to understand their significance within the system.

Institutional repositories

Both e-grey literature and new electronic resources are generally produced by individuals or by organizations where publishing is not the principal activity and therefore they are rarely available through formal channels of communication. This makes them notoriously difficult to access (Auger 1988). For example, reports in the fields of energy or archaeological excavations (Culter 1999; Meckseper and Warwick 2003). As mentioned previously a main driver for repository development has been to make academic research output available to a much larger community and thereby increasing its visibility and impact. Therefore, repositories can serve as useful tool to discover digital resources produced by academics.

Although repositories could be defined as *a collection of digital objects* this term is not useful in the sense that it does not differentiate repositories from other information systems such as databases or catalogues (Heery and Anderson 2005). Repositories can therefore be defined through a series of their characteristics that differentiate them from other collections of digital objects. Heery and Anderson describe a typology of repositories according to functionality, coverage, content types and user group (Heery and Anderson 2005). These will be examined in more detail in the literature review but for the purpose of defining the scope of this research *coverage* and *content type* will be used as tools to qualify the term repository.

The content type refers to the particular characteristics of the material that can be deposited within a repository. Some repositories for example, will only accept a certain type of material, for instance, e-prints or theses and dissertations. In other cases, they will only accept material related to a particular subject area, such as Engineering or Library and Information Science[1]. Coverage refers to the community of users who deposit material. These can be personal (for example an author's personal archive), a journal (depositors are authors or editors), departmental/ institutional/inter-institutional (members of the department or an institution(s) may deposit), or regional, national and international (depositors are from a particular

[1] For example, E-LIS is an e-print repository for papers in the field of Library and Information Science. See http://eprints.rclis.org/

geographical region or open to all). Together the coverage and the content will qualify the type of repository. For example: an *institutional* repository, a *subject* repository, an *e-print* repository, an *institutional thesis* repository, a *national e-print* repository.

For this thesis it was considered that in terms of coverage, institutional repositories belonging to higher educations institutions (HEIs), in particular universities, were an effective way of discovering digital resources produced by academics. University repositories are fairly stable, backed by established institutions and easily delimited. Institutional repositories are also now the most popular type of repository; there are currently over one thousand institutional repositories worldwide, in contrast to less than two hundred subject repositories and less than thirty governmental[2].

In relation to content type, the research scope for this thesis is informal electronic resources. Therefore, repositories that collect material that corresponds to the formal electronic publishing framework, such as pure e-print repositories, were not considered for this work. It was a necessary condition that repositories should contain non-formal resources. However, repositories should also handle a wide variety of informal resources, therefore monotype repositories, such as dataset repositories, are not within the scope either. Repositories should handle a range of different informal electronic resources. It is important to add however, that repositories that contain a wide variety of resources will almost always include both formal and informal electronic resources. Therefore, although the focus is on non-formal resources the institutional repositories contemplated for this study will contain a mixture of formal and informal resources. In summary, for this research, the focus will be on institutional repositories as these contain a wide variety of electronic resources and have an institutionally defined user base.

[2] Data taken from Open DOAR (Directory of Open Access Repositories) November 2008. The OpenDOAR directory and other repository listings are discussed in further detail in the Methodology section.

Thesis outline

The research is divided into five chapters and five appendices. The opening chapter – Introduction – describes the background to the research, the motivations for the study, the aims and objectives of the work and the working hypothesis and limits the scope. The second chapter reviews the literature on institutional repositories, definition of electronic resources, electronic publishing and its implications for scholarly communication and publishing and puts this study in context. The third chapter describes the different methods used for this study: online survey for repository managers and the seven case studies that included link analysis and interviews with repository managers, discussing their characteristics and usefulness for this study. Chapter four presents the results of the different studies, including the variety of repository materials, perceived functions of the repository, and results of the different types of materials that are linked to within the repository and discusses the findings. The fifth chapter presents the conclusions of this research as well as the limitations of the study and the possibilities for further research in this area.

Chapter 2- LITERATURE REVIEW

Conducting a literature review for relevant research on the subjects of repositories (institutional repositories in particular), the definition of electronic resources, the changing role of informal and formal publishing in the networked environment and all within the context of electronic publishing was a challenge in itself. A large amount of the work done in these subjects has theoretically explored the issues but there is little research on the actual instances. As this is a relatively new area of enquiry, one of the main attractions is that there is plenty of scope for research but one of the drawbacks is discovering the relevant literature. This review is not just an analysis of the relevant literature that was collected, but it is also an exercise that attempts to describe the key elements and characteristics of this mainly uncharted territory. This literature review therefore, not only analyses and synthesizes the available previous work on the subject (Levy and Ellis 2006), but also gathers and collectively overviews literature on these issues providing an initial foundation for this new research topic.

Three important challenges arose during the literature review. The first important issue is the inherently multidisciplinary nature of electronic publishing and institutional repositories, in so far that they affect all disciplines across the board. Conducting a literature review on these subjects required searching extensively across different disciplines. Electronic publishing as a research area in itself has a few publishing and communication channels; the most obvious ones being *The Journal of Electronic Publishing*[3] which ran from 1995 to 2002 [4] and the Electronic Publishing Conference which celebrates its 13th anniversary this year. The Scholarly Electronic Publishing Bibliography (Bailey 2006) has been published since 1992 (Bailey 2001) and

[3] See Journal of Electronic Publishing (http://www.press.umich.edu/jep/)

[4] The JEP has recently come back in print with the February 2006 issue. Originally started up by the University of Michigan Press the journal was unfortunately not published for almost four years (2002-2006) missing out on important times for electronic publishing. The Scholarly Publishing Office of the University of Michigan University Library has now assumed publication responsibility as an experiment in library-based publication. For more information see Bonn (2006).

continues to be a source for information on the subject. This is now published together with the Open Access Bibliography (Bailey 2005-2007) under the umbrella of Digital Scholarship.

As expected, the Library and Information Science field has contributed to the fields of electronic publishing and repositories with publications such as *D-Lib Magazine*, *Journal of Documentation*, *Aslib Proceedings, Journal of the American Society for Information Science* and many others, frequently publishing related articles. Information technology, computing systems and computing engineering also have relevant research. Publications on the study of the Internet have also dealt with electronic publication issues as can be seen for example, in *First Monday*.

Other related subject areas are Human Computer Interaction (HCI), Hypertextual studies, Humanities Computing, e-learning and others. For example, journals such as the *Literary and Linguistic Computing* and recently *Digital Humanities Quarterly*, have also published related work. All have discussed electronic resources and publishing in their own context. In recent years, especially with the advent of Open Access (which will be discussed in more detail in this review), electronic publishing discussions have been especially active in the Sciences. For example, a seminal discussion on Open Access was originally published in *Nature* (Okerson and O'Donnell 1995). In addition a number of other scientific journals have published discipline related electronic publishing articles. As would be expected the publishing field has also contributed with journals such as *Learned Publishing* being one of the main channels.

Several national committees have also published important reports covering several electronic publishing and repository issues. In the United States the Research in Digital Libraries Initiative (DLI) has produced related research since 1993. In the United Kingdom the Joint Information Systems Committee (JISC) has produced numerous reports[5] and funded projects

[5] For more information see the JISC website (http://www.jisc.ac.uk/)

such as Focus on Access to Institutional Repositories (FAIR)[6]. This work will be reviewed in this section.

The second important challenge for conducting this literature review has been the lack of consistent terminology in the electronic publishing and repository world. This could be a result of the multidisciplinary nature and relative newness of the subject area. It is likely that each discipline that has contributed to the electronic publishing and repository literature has, probably unknowingly, imported phrases and taken certain concepts for granted that in other areas work differently. In this sense, Kling and McKims' remarks (Kling and McKim 2000) warning against treating all scholarly electronic publishing as the same, should be applied to not treating all literature on electronic publishing as the same. The discipline perspective, at least for the time being, is important.

Another factor that may explain this semantic ambivalence is related to the more complex relationships that are at stake when discussing electronic publications and institutional repositories. From the onset the mere possibility of electronic publishing and repositories as opposed to traditional print publishing directly challenged and questioned the role of each one of the players in scholarly communication and publishing; authors, editors, typesetters, printers, distribution agencies, subscription agents, librarians, teachers, researchers and students (Willis, 1996). As pointed out by Jones over ten years ago, during a colloquium on scholarly communication issues was that what was often heard were declarations and assertions that staked out and fortified a position (Jones 1998) and this is arguably still the case today. This is not to say that there has not been productive and cooperative exchange of ideas but some aspects of electronic publishing and repository literature and discussion has been more of a scramble to assert and hold power positions and less of a rational academic discussion. If we view electronic publishing as a means of empowerment or disempowerment, whether it be

[6] See Focus on Access to Institutional Repositories (http://www.jisc.ac.uk/index.cfm?name=pub_fair)

economic losses or gains, influence and recognition or control –gate keeping- then the absence of common vocabulary is not due to a lack of academic rigour but signs of a major power struggle beneath the surface.

It is important that this be kept in mind when reviewing the literature on repositories and the effects of electronic publishing on scholarly communication and publishing. There are technical issues of course, but the main focus for this research will be on the sociological aspects. Many of the factors that will affect the role of electronic resources within institutional repositories and their possible impact on scholarly communication and publishing are inherently more social and culture than technological. "the premise of Socio Informatics [SI] is that social forces help to shape technology, [and] to understand this dynamic requires a discussion of the major social forces involved. The social forces represent multiple perspectives and rarely have clear cut answers" (Kling, Rosenbaum et al. 2005:97).

Socio informatics is a useful methodological tool for approaching, in particular contentious issues, within the literature on repositories and electronic publishing. "The concepts of SI imply a dynamic tension between the positive and negative effects of new ICTs (…) develop an ability to think critically about the roles and values of ICTs (…) examine ICTs from perspectives that do not automatically and often implicitly adopt the goals and beliefs of the group that commission, design, or implement specific ICTs" (Kling, Rosenbaum et al. 2005:96). Repositories for example, have been at the centre of controversial scholarly movements and it could be argued that some papers on the subject describing the merits or the dangers are pushing more of a political agenda than pursuing a research question. In this sense what is currently lacking is an evaluative framework to be able to determine how to measure the success of repositories. The aim of this literature review is to analyse the literature but also to provide a useful theoretical framework in which to view repositories and their role in electronic publishing more objectively.

The literature review is divided into four sections: the first section, examines the definition of electronic resources and the literature on the subject. In particular the difficulties encountered for defining non-formal e-resources. The second section looks at electronic resources in relation to institutional repositories[7], which are becoming a popular solution for the academic community to manage and disseminate disparate electronic resources. The origins of IRs are reviewed, and in particular the studies on the key drivers behind their creation. The perceived functionality of IRs are explored in relation to studies on the different types of electronic resources that are contained with them. Electronic resources are particularly difficult to define and to detect and this section also gives an overview of a range of methods and studies that have been developed in order to attempt to effectively measure the use of these resources and to understand what they are being used for. The third section, examines the literature on electronic publishing regarding concepts of 'publishing' and its relationship to scholarly communication in general, and its implications for institutional repositories in particular. The final section, examines the literature on the future of electronic publishing, in particular with relation to the role of electronic resources and institutional repositories.

Electronic resources

A research resource is a generic term that refers to a broad range of materials that can be useful for researchers. In the same sense an electronic resource can mean anything from the journals available electronically at a university library to a list of web links (Spark Jones, Bennett et al. 2005). Few studies were found that attempt to define the term electronic resources in sharp contrast to the much larger body of work on the lack of consistent terminology in the electronic publishing world. This provides further evidence for the need to develop more sophisticated digital genre typologies. In this section a selection of studies on electronic resources are revised in order to see how the term can be applied to this thesis, followed by a revision of other phrases, such as grey literature, in order to put non-formal e-resources into context.

[7] Institutional repositories henceforth referred to as IRs

Defining electronic sources

In their report the British Academy (Spark Jones, Bennett et al. 2005) acknowledges the difficulties involved in defining electronic resources, as these in principle could be anything available in an electronic format. However, they make a useful differentiation between primary and secondary electronic resources, as a way of approaching a definition. They make a broad distinction "between resources *on* which research is done, i.e. primary resources, and resources *through* which primary resources are reached, i.e. secondary resources" (Spark Jones, Bennett et al. 2005:10). Examples of primary and secondary resources are summarized in Table 1.

PRIMARY ELECTRONIC RESOURCES	SECONDARY ELECTRONIC RESOURCES
Field notebooks, manuscripts, working drafts, theses, conference proceedings, reports, graphics, maps, photographs, satellite images, numerical data, text corpora, journals and books	Library, archive and museum catalogues, bibliographies, web portals, search engines, abstract journals.

Table 1. Primary and secondary sources

The distinction between primary and secondary electronic resources or discovery tools is a useful one as both types of resources tend to be bundled together although their nature, use and characteristics are very different. Warwick (Warwick, Terras et al. 2006) found that although scholars defined themselves as users of digital resources, when asked to list their three favourites these were actually generic information resources, i.e. secondary resources, such as Google, Humbul, Web of Knowledge, as well as the university library. This confusion or ambivalence is important when doing research into the use of electronic resources, as primary or secondary resources will have very different characteristics and purposes. For many, electronic resources are a means of accessing information (like an archive or a library), rather than the object of study in itself (for example a digital monograph) (Warwick, Terras et al. 2006).

The confusion between primary resources and secondary resources is not surprising, considering the common ambivalence of terms in electronic publishing in general. Although electronic publishing has been widely researched and discussed, little attention has been paid to the terms used. This insistence on lack of discursive consistency is important and can be illuminating when attempting to wade through the 'cacophonous discussion' (Kling and McKim 1999:890) regarding electronic publishing and resources. Much of the published research uses interchangeably or lightly terms such as e-publishing, e-journals, e-prints, e-manuscripts, post prints, electronic resources and other terms such as putting online, posting online and publishing online. For digital books and the readers (e-books, e-book reading appliance, electronic books, e-book readers, among others) the situation is similar with imprecise and inconsistent terminology (Lynch 2001). Although the terms are familiar there appears to be an absence of a common vocabulary. Many of the ensuing discussions with regard to the future of scholarly publishing utilize these terms and the semantic instability (Cronin 2003) hampers productive discourse.

The distinction between primary and secondary resources serves as a starting point for this thesis by regarding items within the repositories as primary electronic resources and the repositories themselves as secondary electronic resources. The term electronic resources throughout this thesis will be used to refer to primary resources.

It is important to note that the British Academy report (Spark Jones, Bennett et al. 2005) makes no distinction between published and unpublished primary electronic resources. Although this is not a criticism of the report, it is a crucial distinction for this thesis, in particular when observing the implications of unpublished electronic resources on more traditional conceptions of scholarly publication and publishing. The concept of publishing for electronic resources is examined in detail in a later section of this chapter.

Meadows (1998) makes a similar distinction between primary and secondary resources when using literature for research. He differentiates between primary (journal, articles, books) and secondary literature (abstracts, journals, indexes and so forth) with the latter serving as a tool to discover the former. He notes that it was actually secondary literature, such as indexes and databases, that was first made available in the electronic environment, with the full text being added afterwards.

Again, this thesis is focusing on primary literature with institutional repositories acting as a form of secondary 'literature', as an aid and discovery tool of the primary resources. However, the scope of primary literature is much broader than the formal examples offered by Meadows. He does point out though that one of the differences between the printed and the electronic world is that there is an increased blurring of the distinction between data, information and knowledge. The print world has more specific channels for conveying these different types, whilst in the electronic world raw and refined information can become mixed. Within institutional repositories the range of electronic resources will mean that they may harbour raw data (for example, automated measurements) with more thoroughly analysed and discussed resources such as a report or a preprint. How can these be handled? What are the implications in terms of use? Will it be important to differentiate between them and how will this be done? A key aspect towards answering these questions is to understand the electronic resources that are not within the formal realm and therefore fairly standardized and identifiable, but rather the ones that we have referred to as non-formal e-resources.

Defining non-formal e-resources

There appears to be no set name for academic electronic resources that are outside the framework of formal electronic publishing. However, the existence and use of materials that are outside of the formal framework of scholarly publishing is not new. In the print world, these materials are generally referred to as "grey literature". This term is a good starting point, as some electronic resources tend to share many grey literature characteristics, both in the type of

material as well as matters related to their access, reliability and discovery. For example, it has been pointed out that some synonyms for the word 'grey' used in the professional press are 'non-conventional', 'informal', 'informally published', 'fugitive' and even 'invisible'. (Auger 1988). These terms are similar to those applied to some types of electronic resources, in particular newer ones such as blogs, wikis and web pages.

Grey Literature Network Service define grey literature as: "Information produced on all levels of governments, academics, business and industry in electronic and print formats not controlled by commercial publishing, i.e. where publishing is not the primary activity of the producing body" (GreyNet 2004). It is also "available through specialized channels and may not enter the normal channels or systems of publication, distribution, bibliographic control, or acquisition by booksellers or subscription agents" (GreyNet 2004). Electronic resources are generally put online by the creators themselves usually scholars or academic groups. Although in theory these resources are now available to anyone using the Internet, the discovery of good quality, reliable and useful electronic resources is a major problem (Greenstein and Trant 1996; JISC 2005; Dunning 2006).

Institutional repositories are not controlled by commercial publishing but rather by academic institutions themselves, usually through the Library. They collect academic materials, provide access and dissemination to them and ensure quality to varying degrees. It could be argued that IRs are partially fulfilling some traditional publishing roles. They are currently not considered a 'normal' channel for academic communication. However, this is something that arguably could change. For example, arXiV albeit a subject repository, is now considered among the Physics community as an accepted channel for communication (Aymar 2009; Mele 2009). Issues surrounding the roles of publishers and publishing are examined in more detail in the section *What is publishing? A few proposals*. An important area of study is to analyse the role of IRs in relationship to publishing and to examine possible ways that these two mechanisms can interact

and complement each other. In this sense it would be pertinent to ask whether IRs, with time, will become a 'normal' channel for communication?

Other defining characteristics of grey literature are lack of standardisation and a wide variety of different formats: "Precisely because grey literature is so amorphous and intended for a wide variety of purposes, it is not obliged to confirm to the standards of presentation imposed by the editors and publishers of conventional publications" (Auger, C. 1988:2). Important issues are how to present, manage, store, cope with versioning, distribute and ultimately preserve the wide variety of electronic resources that are being created and deposited in repositories. There has been some work in attempting to standardise electronic resources. For example, the AHDS (Arts and Humanities Data Service) for the UK, ICT (Information and Communication Technology) Guides for Arts and Humanities research have contributed towards establishing standards and best practices for the creation of Humanities digital resources (Spark Jones, Bennett et al. 2005). However this is, generally speaking, still a pending issue in particular for the unstructured environment of non-formal e-resources. An indispensable first step to understanding this landscape is to have a more detailed look at the range and types of electronic resources.

On the other hand the web, as a medium for dissemination and publishing, offers a new range of possibilities for grey literature (Banks 2005; Lambert, Matthews et al. 2005). The Internet can also be seen as a low cost, easy entry-level medium for publishing material online. Successful examples of blogs, websites, databases, videos, digital collections, among others have shown the amazing capacity of the web as a medium for attracting a large number of users, and bypassing the traditional actors of publication and distribution. Examples of using new types of resources for research (and teaching and learning) are: the British Library's Turning the Pages

Project, the Human Genome Project with the Gene Gateway, GBif (Global Biodiversity Information Facility) for biological collections and Project Gutenberg, to name a few[8].

In summary the grey literature characteristics that can be useful for providing insight into the realm of non-formal e-resources are:

- no fixed standards of presentation
- not controlled by commercial publishers
- publishing is not the principal activity of the producing body
- not available through the 'normal' channels of distribution

It will be particularly interesting to note what can occur with non-formal e-resources if they are stored within IRs. Will IRs take over certain roles performed by publishers? Are they effectively becoming publishers? Will they contribute towards improving standards of presentation? Will IRs become an effective new channel of access and distribution? Will publishing formal and/or non-formal e-resources become a new role for the university library?

In the networked environment grey literature can also encompass information produced in a specific working context which is, or might be of value outside that context (Lambert, Matthews et al. 2005). In this sense, material that is produced for a specific purpose, such as a final dataset, can serve, for example, as a baseline data for another project. Along these same lines, there is an increasing emphasis on creating electronic resources that can be repurposed for the needs and objectives of other scholars involved in different activities. For example, the reports for the commission of Cyberinfrastructure (NSF 2006), emphasize the need for electronic resources and datasets, that can also be manipulated, reused and repurposed both in Science and

[8] For more information see:
BL Turning the pages: http://www.bl.uk/onlinegallery/ttp/ttpbooks.html
Human Genome Project: http://www.ornl.gov/sci/techresources/Human_Genome/home.shtml
GBif: http://www.gbif.org/
Project Gutenberg: http://www.gutenberg.org/

Engineering (Atkins, Droegemeier et al. 2003) and Social Science and the Humanities (Hey and Trefethen 2005). The NSF (National Science Foundation) and the JISC (Joint Information Systems Committee) joint report on the future of scholarly communication (Arms and Larsen 2007) also argue for a digital content infrastructure that will allow new approaches to scholarly research that are collectively referred to as cyberscholarship.

Moreover, as different types of material become available online and users discover material through alternate channels, there will be an increased blurring of formal and informal publishing. Banks (Banks 2005) suggests a *continuum of scholarship* in which there is a collapse of the distinction between grey and non-grey literature. This concept is not unlike Kling's *continuum of publishing* (Kling and McKim 1999), which will be discussed in a later section. However, precisely because of this predicted blurring, e-grey literature is not helpful as a term to describe non-formal electronic resources, despite their similarities. Nonetheless, it is useful towards developing an approach to define the characteristics of electronic resources that are outside the framework of formal electronic publishing.

As mentioned in the scope section, the emphasis of this study is on electronic resources that are outside the framework of formal electronic publishing. There is however, a caveat in relation to electronic e-prints. These can be divided into two types: preprints and postprints. Preprints are "manuscripts that have not yet been published, but may have been reviewed and accepted; submitted for publication; or intended for publication and being circulated for comment" (n.d. US Department of Energy, Office of Scientific and Technical information quoted in McKiernan 2000:127). Postprints on the other hand, are the final, refereed version of the article (Harnard 2001) and belong the formal publishing world. In the print world, preprints were generally considered grey literature although their final destination was the formal electronic publishing world. In the online world however, preprints within repositories are usually linked to the formal electronic publishing. This is probably due to the fact that the intention is that they will later be replaced with the postprint. It could be argued however, that they still share many

characteristics of informal or grey literature. For example, some of them may not have yet been peer reviewed or accepted for publication. Given that preprints are an important constituent of institutional repositories and although the focus of this research will be on non-formal electronic resources, preprints must be mentioned separately. As discussed, divisions are not always clear-cut and the distinctions between formal and informal e-resources will not always be decisive. This will be accounted for in the research.

The next section examines institutional repositories, the secondary electronic resources, that will provide the framework for studying electronic resources and in particular non-formal e-resources.

Institutional repositories and electronic resources

Overview

Institutional repositories first appeared in 2002 as an institutional response to the increasing trend for scholars to post their research online, usually on their homepages (Johnson 2002) but also in subject based repositories. Repositories are associated with a number of different scholarly initiatives and there is a large body of literature that describes IRs and explores their role within scholarly communication and publishing. The Scholarly Electronic Publishing Bibliography (Bailey 2006) contains over 200 entries on the section *Repositories, E-Prints and OAI*, whilst the Open Access Bibliography (Bailey 2005-2007) also contains over 200 entries in the *Institutional Archives and Repositories* subsection. The following section will provide an overview of the main drivers behind the creation of institutional repositories, examine the current literature on institutional repositories and how the term will be used within this work.

Development of institutional repositories

This section provides a brief overview of the history of repositories, including preprint servers and the development of enabling technologies such as OAI-PMH (Open Archives Initiative – Protocol for Metadata Harvesting) and open source repository software. We look briefly at the

Open Access movement in the context of institutional repositories. Seminal papers on IRs are examined in detail with particular focus on the different drivers behind the creation of institutional repositories together with specific emphasis on the implications for the types of materials that are collected.

Online archives -The history of digital preprints and postprints

The origins of institutional repositories can be traced back to the early 1990s when the new, networked academic environment allowed researchers to communicate faster using the new digital communication channels. In certain subject areas, such as Physics and Mathematics, scientists were accustomed to distributing preprints -versions of articles that had been submitted to journals but were still not published- among interested colleagues. An informal way of communication (McKiernan, 2000) it helped researchers keep up to date with the most recent research as the time lag between journal article submission and publication could be quite slow. Harnard suggested that one of the biggest impacts of the revolution of electronic networks could be on prepublication: "On the brink of intellectual perestroika is the vast PREPUBLICATION phase of scientific inquiry in which ideas and findings are discussed informally with colleagues (currently in person, by phone and by regular mail), presented more formally in seminars, conferences and symposia, and distributed more widely in the form of preprints and tech reports" (Harnard, 1990: 342). Using networked technologies, such as FTP and email, preprints could be distributed to fellow academics around the world almost instantaneously and at a low cost. By placing preprint version of articles on a central server, other academics would be able to download the full text of the latest research.

Another crucial distinction with preprints in the print world is that the digital archives would allow, during this informal prepublication stage, immediate feedback from peers by permitting them to comment on the preprint. This open peer review system would enrich the article and allow the research process to be more open. Harnard (1990) describes the possible mechanisms that would facilitate what he called Scholarly Skywriting, allowing scholars to submit preprints

of their manuscripts to a global community for immediate feedback. What is important to note is that Harnard is proposing a *prepublication continuum of scientific enquiry*, that in many senses is similar to Banks (2005) *continuum of scholarship* and Kling and McKim's (1999) *continuum of publishing*. One key distinction though is that Harnard's continuum is limited to prepublication as he continues to uphold the need for the formal publication process. The continuum is within the prepublication process but there is no blurring between published research in formal and informal channels.

Originally known as the LANL (Los Alamos National Laboratory) preprint archive, and later renamed arXiV, this was the first online archive for preprints in Physics and later expanded to other subject areas (Astronomy, Mathematics, Computer Science, Nonlinear Science, Quantitative Biology and Statistics). Created by Paul Ginsparg in 1991 the archive originally used email to alert subscribers to new submissions who could then access the full text on demand using FTP server technology (McKiernan 2000; Ginsparg 1996). It has been and continues to be used as an example of a successful archive currently containing over half a million articles and heavily used (Aymar 2009; Mele 2009).

What is particularly important to note for this thesis, is that these online archives were a way for scholars to communicate and share their articles in their preprint version. Both Ginsparg (1996) and Harnard (1990) were clear that these archives would serve well for royalty free literature or what Harnard referred to as esoteric scholarly publication (Okerson, 1994). These are publications that have been written solely for research impact and where there is no expectation of royalties (Harnard, 2001). Academics write and publish articles in order to communicate their results to their interested audience (which is generally rather small and specialized) and for recognition in order to advance their careers.

It was this line of thought that led to the 1994 discussion *A Subversive Proposal* (Okerson, 1995), an email discussion that took place over a period of months, in which Harnard proposed

that the networked environment, through the use of online archives would allow authors to break the *Faustian bargain* that they entered with publishers. Subscription barriers set by publishers were limiting rather than enhancing the access to the articles and academics needed to find new ways to increase dissemination and access mechanism in order for their research to reach their audience. Harnard and others (Okerson, 1994) proposed that e-print archives could be used not only for preprints but also to distribute postprints, i.e. the final refereed version of the article. Once an article had been published then the preprint could be replaced with the final refereed postprint version. In this fashion all refereed scientific *esoteric* academic literature could be available with no barriers[9].

It is particularly worth noting that up to this point the development of e-print archives mainly focused on improving distribution and access to journal articles, either as preprints or as postprints, rather than disrupting or changing the scholarly communication publishing system in general. It was believed that *communication* of research results in the form of *esoteric publications (i.e. journal articles)* could be improved using online e-print archives.

Creation of OAI-PMH

The arXiv preprint server became a key example of the possibilities for dissemination of scholarly material. The advent of the World Wide Web solved many of the searching, navigation and retrieval difficulties of the FTP server (Harnard, 2001). As other e-print archives were developed, such as RePEc (Research Papers in Economics[10]) and CogPrints (Cognitive Science Eprints Archive[11]) (Van de Sompel and Lagoze 2000) , issues related to interoperability between servers and cross searching were raised. It became important to be able to exchange information between archives in order to make the e-prints more accessible and searchable. In

[9] The remaining barriers at this point would of course be technological, such as access to the Internet, a computer and so forth but they would no longer be related to journal subscription fees.

[10] Founded in 1993. See http://repec.org/ for more information.

[11] Founded in 1997. See http://cogprints.org/ for more information.

1999 a meeting in Santa Fe led to the establishment of the Open Archives Initiative (OAI)[12]. The Santa Fe Convention is a "combination of organizational principles and technical specifications to facilitate a minimal but potentially highly functional level of interoperability among scholarly e-print archives" (Van de Sompel and Lagoze 2000:unpaginated). The convention wanted to create functional but easy to implement recommendations in order to ensure a low barrier entry for all e-print archives. Recommendations were therefore kept at the metadata level.

The OAI eventually led to the development of the OAI-PMH (Protocol for Metadata Harvesting). The technical specifications of the protocol are largely out of the scope for this literature review but two crucial elements must be noted as they have important repercussions for the future development of institutional repositories: the separation of metadata and digital object and the choice of Dublin Core metadata as standard.

OAI-PMH recommends the use of the Dublin Core, a metadata schema for resource description, although other schemas such as MARC may be added. Dublin Core is deliberately simple and provides a standarised metadata set of 15 elements and is widely used to describe digital resources online[13]. Dublin Core was selected precisely because of its simplicity as it could easily be filled out by academic themselves when they self-archived their e-prints. As repositories have moved away from e-prints toward storing more complex non-textual objects, in particular multimedia resources, severe limitations have been found in practice with using Dublin Core to describe these resources or attempting to append additional metadata schemes (Emly 2007). As will be discussed later, this is an important consideration when discussing non-formal electronic resources within institutional repositories.

[12] This was originally called the Universal Preprint Service, but this was changed due to the similarity between its acronym UPS to the US international courier service. What is important to note is the importance of the preprint as the format.

[13] For more information see the Dublin Core Metadata Initiative website (http://dublincore.org/)

The protocol effectively separates the metadata from the digital object (in this case the e-print). The idea behind this was to allow archives to act as data providers, providing their metadata for harvesting by service providers. The metadata *record* may or may not have the full text or digital object associated with it. Although it was always the intention for the digital object to reside within the repository, it is technically possible to implement OAI-PMH with metadata only records and no digital objects attached. As will be discussed later, this has led to institutional repositories containing records but not necessarily the digital objects they are referring to. In terms of the role of institutional repositories as providers of *access* to electronic resources this naturally has implications that will be examined.

Software creation

Another outcome of the Santa Fe convention was the recognition of the importance of developing software so that other institutions could begin to create their own archives (Van de Sompel and Lagoze 2000). Based on CogPrints (Tansley and Harnard 2000) developed at the University of Southampton, work began on creating an easy to install software that implemented OAI standards and allowed the self-archiving of e-prints. According to their specifications archives should have:

- A submission mechanism
- A long-term storage system
- A management policy with regard to their submission of documents and their preservation
- An open machine interface, that enables third parties to collect data from the archive (effectively OAI-PMH) (Van de Sompel and Lagoze 2000)

In 2000 the appropriately named EPrints was launched by the University of Southampton, followed a couple of years later by DSpace produced by the MIT in conjunction with Hewlett Packard (Smith, Barton et al. 2003). This was followed by other software such as Digital Commons offered by BePress and the use of tools such as Fedora and Greenstone for

institutional repositories[14]. Figure 1 shows usage of repository software in repositories registered in OpenDOAR (Directory of Open Access Repositories)[15] from November 2008. EPrints and DSpace are considered to be the leading software for repository development.

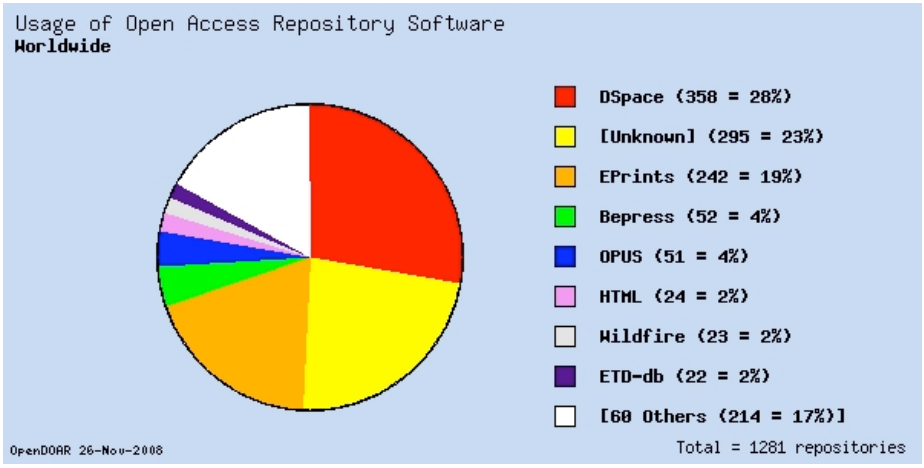


Figure 1- Usage of Open Access repository software worldwide

It has been suggested that the choice of software for the repository can have repercussions on the types of materials that are collected and the ways in which they are handled, although this has not been thoroughly studied. The DSpace information model is based around the idea of Communities that manage Collections. DSpace allows different Communities to set their own collection policies, including permissions to deposit, types of materials that are allowed for deposit and so forth. The software uses a system for persistent identifiers in order to help ensure long-term stable access and aid with preservation issues (Smith, Barton et al. 2003).

EPrints on the other hand, was developed from the e-print server technologies and is developed to be an out of the box system (Tansley and Harnard 2000). In an analysis of different

[14] Both DSpace and ePrints are open source software, whilst BePress, through Digital Commons, is a licensed service offered by the Berkley Electronic Press. Greenstone is better known as a digital library tool that has been adapted for institutional repositories. Fedora is an open source digital asset management (DAM) tool upon which different types of digital libraries, repositories and archives can be built. More detailed information can be found:

DSpace: http://www.dspace.org/
EPrints: http://www.eprints.org/
BePress (Digital Commons): http://www.bepress.com/ir/
Greenstone: http://www.greenstone.org/
Fedora Commons: http://www.fedora-commons.org/

[15] A comprehensive description of the OpenDOAR service is available in the Methodology section.

repository software, Sale (2005) indicates that EPrints requires little technical expertise to install but that larger universities will possibly require more powerful software options such as DSpace or Fedora. In general is has been argued that the current software is adequate and it is technically relatively easy to set up a repository. This seems to be the case particularly for plain e-print repositories. However, as repositories attempt to incorporate more complex non-textual materials a few studies have documented and indicated difficulties when attempting to use this software for non-formal resources (Emly 2007; Salo 2008; Shreeves and Cragin 2008). This being said however, all software is in constant development and the recently launched version of Eprints 3.0 seems to have expanded its functionality to more effectively embrace non-formal e-resources. They now describe themselves as a solution to set up repositories with "research literature, scientific data, student theses, project reports, multimedia artefacts, teaching materials, scholarly collections, digitised records, exhibitions and performances"[16]. What is important to note from this brief analysis of repository software development is that there has been a progression from more eprint orientated functionality towards embracing a wider array of digital materials, although whether this has been successful is still under debate.

Open Access

With the development of OAI-PMH and software for creating archives the possibilities of increasing access and dissemination of e-prints became more of a reality. If academics around the world archived their pre or post print articles in repositories this would effectively eliminate access barriers to scientific literature and increase visibility and access.

The Budapest OA Initiative 2001 states that by Open Access they mean: "free availability on the public internet, permitting any users 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" (Chan, Cuplinskas et al. 2002).

[16] From EPrints website : http://www.eprints.org/ (accessed 26/11/2008)

What is most important to point out is that it was focused on *world-wide electronic distribution of the peer-reviewed journal literature*, although they do mention preprints. The main idea was for the preprint to be substituted with the postprint once this had been published. This statement refers to research output that is still very much within the realm of formal publishing.

The Bethesda Statement on OA Publishing 2003, also referred to primary scientific literature. As OA became more important it became crucial to assure that the archive, now also called repository, was backed up effectively in some way in order that access to the digital resource could be considered permanent: "(at) least one online repository that is supported by an academic institution, scholarly society, government agency, or other well-established organization that seeks to enable open access, unrestricted distribution, interoperability, and long-term archiving" (Brown, Cabell et al. 2003). Issues about preservation, stewardship and long-term management were introduced.

Institutional e-print archives for self-archiving

It was proposed that a main path towards achieving Open Access would be through author self-archiving. In Harnard's paper *The Self Archiving Initiative* there appears to be one of the first mentions of an institutional based archive: "authors need only deposit their refereed articles in e-print archives at their own institutions; these interoperable archives can then all be harvested into a global virtual archive, its full contents freely searchable and accessible online by everyone" (Harnard 2001:1024). As the Open Access movement gained momentum the lack of subject archives encouraged the creation of institutional archives where academics could deposit their e-prints. The OAI-PMH would allow service providers to cross search these archives. The future of e-prints archives and the Open Access movement lay in the creation of institutional repositories for academics to self-archive their published articles (or preprints to later be replaced). And who better to manage these institutional archives than the Libraries themselves who had experience in collecting, managing, cataloguing and preserving?

Institutional repositories and the reformation of scholarly publishing

The increased involvement of libraries in the development of institutional repositories provided another fundamental shift. For Harnard and others the e-print servers were anchored in the need for more rapid communication and increasing visibility (referred to as an 'eyeball issue') in order to assure maximum research impact. For libraries, struggling with increased subscription fees, the movement promised to alleviate their budget woes whilst still providing university members access to research. The background to this is the *serial pricing crisis* (King and Tenopir 1999; Houghton, Steele et al. 2004). First noted in the early seventies (King and Tenopir 1999; Guédon 2001b) by the early 90's it was a widespread concern, in particular within the library community. Libraries argued that during the second half of the twentieth century there was a rapid increase in journal prices well above the inflation rate (Houghton 2001) that was not justifiable in any sense. Commercial journal publishers denied this and argued that there was a justifiable increase due to inflation, currency exchange rates and in order to support new, less lucrative journal titles in more specialized disciplines (Guédon 2001a). Libraries saw in the Open Access movement the potential to alleviate or even resolve this crisis.

In 1998 the ARL (Association for Research Libraries) founded SPARC (Scholarly Publishing and Academic Resources Coalition) to address the serial crisis and correct the imbalances in the scholarly publishing system (Joseph 2006). It is important to mention the creation of SPARC as it has had a major role in the support of the development of Open Access and impulse behind the development of institutional repositories. The SPARC position paper *The Case for Institutional Repositories* (Crow 2002) was a key development in the history of institutional repositories. As mentioned previously libraries saw the opportunity with the Open Access movement to solve or at least alleviate the problem of *access* to published literature that had become a pressing problem due to the increasing prices of journals. Harnard had argued twelve years earlier (Harnard 1990) for a revolution in prepublication, people were now discussing a revolution in the scholarly publishing system as a whole.

Institutional repositories were no longer conceived just as a tool for providing faster access and broader dissemination but positioned as a critical component in reforming the system of scholarly communication. Institutional repositories "offer a strategic response to systematic problems in the existing scholarly journal system". The SPARC paper suggests that a new publishing paradigm can be created in which the different aspects of the scholarly publishing model can be disaggregated. This would allow more flexibility and room for manoeuvre to establish a better publishing system (Crow 2002:29).

Additionally the institutional repository would allow an individual institution to collect and showcase their research. Currently a university's research output is diffused over thousands of journals but if each individual researcher self-archived their publications in their institution's repository, then universities would be able to collect all e-prints in one place. This would then serve as a "tangible indicator's of a university's quality and to demonstrate the scientific, social, and economic relevance of its research activities and would ultimately increase the institution's visibility, status, and public value" (Crow 2002:37).

Although the position paper is focused on changing the scholarly publishing system it does not propose any changes in the types of formats for communicating research and still stays focused on journal articles. It does refer rather vaguely to "intellectual output" but closer reading indicates that the main focus is on formal intellectual output and in particular journal articles. The paper however, does see institutional repositories in conjunction with complementary digital repositories. It is probably these that they are referring to when they say "[the disaggregated] model includes not only preprints and research papers, but also extends to research data sets, digital monographs, theses and dissertations, conference papers, listserv archives, and other gray literature" (Crow 2002:12).

Institutional repositories as infrastructure

Although institutional repositories were originally set up to manage and disseminate e-prints (and contribute towards the Open Access movement) it became clear that there was a wider array of electronic research resources that could benefit from institutional repository technology. In his paper *Essential Infrastructure for Scholarship in the Digital Age*, Lynch argues that a growing amount of electronic resources is being produced by academics (Lynch 2003). He notes however, that there is no cohesive system to aid scholars in managing, disseminating and preserving them. An unprecedented burden is put on scholars who create electronic resources as they are usually responsible not only for creating the resource but also for maintaining the website where their resources are located, as well as attempting to incorporate search and retrieval services, metadata and others, in order to aid the user in discovering and accessing the resource. In addition, Lynch points out the potential dangers for preservation and permanence when this responsibility is left to the scholar who created the resource.

In this sense institutional repositories could provide a service, helping academics manage their digital resources. "At the most basic and fundamental level, an IR is a recognition that the intellectual life and scholarship of our universities will increasingly be represented, documented and shared in digital form, and that a primary responsibility of our universities is to exercise stewardship over these riches: both to make them available and to preserve them" (Lynch 2003:328). Institutional repositories are tools to help scholars organize, maintain, administer and disseminate electronic resources that they are currently producing.

This is a key shift in the kind of materials that may be contained within institutional repositories, moving from e-prints (either as pre or postprints) towards a much broader array of materials that are created by university members. Lynch effectively moves IRs beyond just the realm of publishing and places them as a strategy that will accelerate changes in scholarly communication and publishing allowing universities to play a more active role in modernizing

scholarly publishing. However, Lynch is not interested in changing the current publishing system but rather complementing and enhancing it. He argues that viewing IRs as simply a means towards reforming the economics of the current scholarly publishing system is underestimating their potential, as IRs can have a greater impact by providing the infrastructure for a much broader range of new types of scholarly communication vehicles.

Overview of key initiatives

The following section briefly overviews some of the key institutional repository initiatives and projects worldwide. This overview does not intend to be exhaustive as globally there are currently many repository development projects. The main aim is provide an overview of some of the major initiatives, in particular those that have led the way and served as an example for other repository projects.

In the UK, JISC (Joint Information Systems Consortium) launched, in 2002, the FAIR project (Focus on Access to Institutional Resources) to "investigate the technical, organisational and cultural processes involved in providing access to institutional digital resources" (Awre and Baldwin 2006:5). The report concluded that the main challenges for institutional repository deployment were not technical and listed the following as the biggest issues to be addressed by developers:

- Clarity of purpose
- Quality control
- Metadata and semantics
- Legal, ethical and cultural issues
- Research cultures
- Variation between disciplines in terms of methodologies and practices

Under the FAIR umbrella, fourteen repository related projects were run. One group was focused on e-prints and theses and the most relevant projects: ROMEO, SHERPA and TarDis are summarized below:

- TarDis (Targeting Academic Research for Deposit and Disclosure) investigated the social and technical issues for setting up a repository. In particular it enhanced the EPrints software package and addressed issues related to metadata and multidisciplinary requirements for institutional repositories. The project was led by the University of Southampton and led to the creation of the institutional repository e-Prints Soton.

- SHERPA (Securing a Hybrid Environment for Research, Preservation and Access) installed twenty IRs around the UK, and aimed to develop and promote a favourable environment for the UK's research output (namely e-prints) to be made available. One of the main outcomes was the production of a large body of repository advocacy material that has been widely used by start up repository projects. Currently, as SHERPA Plus, the project is aiming to provide a proactive national information point on IRs, offering seminars, courses, email discussion lists, among others. SHERPA also currently runs the RoMEO project described below, and the OpenDOAR registry of repositories described in the Methodology section.

- The RoMEO project (Rights Metadata for Open Archiving) focused mainly on copyright issues related to self archiving eprints. The main aim was to understand stakeholders needs in relation to intellectual property issues that arise with self-archiving and OAI-PMH. An important product was the creation of a database of copyright policies from commercial publishers. This product, know as RoMEO, is now run by SHERPA and is a valuable tool for repository managers, summarizing the permissions that are normally given as part of each publisher's copyright transfer agreement.

FAIR ran an additional three projects under the project cluster Museum Collection and Images focusing in particular on non-textual items such as images and museum objects in order to assess how the Dublin Core metadata standard, recommended by OAI-PMH, could be applied to these types of resources. The project results indicate that although Dublin Core is particularly useful for eprints there are important limitations when used for images and museum objects. It was concluded that further work was required in this area.

In the Netherlands, ARNO (Academic Research in the Netherlands Online) was a project that ran from 2000 to 2002 in order to design and implement digital archives to preserve university output. Between 2003 and 2006, DARE (Digital Academic Repositories) worked to coordinate repository development on a national scale and linked all thirteen Dutch universities and three major academic institutions to form DAREnet. The Netherlands currently has a 100% coverage rate of institutional repositories as all HEIs participated. Their main focus has been in collecting journal literature. Although in 2004 they noted that: "the content mainly consists of text documents, but some photographs and videos are also included. Although the repositories can contain digital objects of any kind, this is yet to be the case in practice" (van der Kuil and Feijen 2004:unpaginated).

On the other side of the Atlantic, the USA currently has 325 institutional repositories[17] nationwide and in terms of simply numbers more than any other country in the world. The development of DSpace repository software was done at the MIT and this in itself has been a major contribution to the institutional repository landscape.

The MIRACLE (Making Institutional Repositories A Collaborative Learning Environment) Project is funded by the IMLS (Institute of Museum and Library Services) and investigates the development of institutional repositories in colleges and universities to identify models and best

[17] Data taken from OpenDOAR on 15th December 2008

practices in the administration, technical infrastructure, and access to repository collections. This project began in 2005 and was originally scheduled to finish in 2008. The results of this census of IRs in the US (Rieh, Markey et al. 2007) are reviewed in a subsequent section *Content and item types in IRs*. Their project plan indicates that they are currently involved in investigating how people search, retrieve, and use institutional repository resources through an analysis of transaction logs and experimental search test tasks. However, this stage of the research project is not scheduled to be finished until August 2009, so the results cannot be reviewed.

Australia is another country with important national institutional repository projects. The ARROW project (Australian Research Repositories Online to the World) identified and tested software solutions for building repositories to handle eprints, electronic theses and dissertations, e-research and electronic publishing. The project also developed and now offers the Arrow Discovery service that uses harvested metadata from all the ARROW repositories.

Summary of repository and IRs development

The previous section has provided an overview of the history of repositories and enabling technologies. In particular we have focused on the shifting key objectives leading from eprint archives to institutional repositories. The following list presents a summary of the main objectives of repositories that have been discussed in the overview of the development of institutional repositories.

Self-archiving and Open Access

- Pre and postprints to increase visibility of esoteric literature (ie. research articles)
- Revolutionize prepublication (rapid dissemination and more peer review)
- Maintain publishing system in place (peer review and certification of research from journal publishers)

(Harnard 1990; Ginsparg 1996; Harnard 2001)

IRs for revolutionizing scholarly publishing

- Digital collections capturing and preserving the intellectual output of a single of multi-university community
- Critical component in reforming the system of scholarly communication (breaking away from monopoly of journals and pricing system)
- Indicators of university's output: increasing visibility, status and public value
- Intellectual output undefined but refers mainly to eprints

(Crow 2002)

IRs as digital infrastructure for universities

- A set of services for members of a university
- Management, dissemination and preservation of digital materials created by them
- Increased breadth of digital materials
- IRs are about scholarly communication, broader than just scholarly publishing
- Support new forms of scholarly communication

(Lynch 2003).

The summary shows how the characteristics and objectives of repositories have been modified over time. Defining an institutional repository is a challenge and the history of IRs helps to shed light on the varying approaches and definitions that are used in the current development and study of repositories.

In order to reach a working definition of institutional repository for this thesis in the next section we shall focus more specifically on current repository functionality in relation to content studies, user groups and usage.

Definition of repositories and institutional repositories

As mentioned in the scope section of this work Heerey and Anderson developed a typology of repositories according to coverage, content type, functionality and user group (Heery and Anderson 2005). In the scope for this study coverage was defined as institutional and the content type as broad, with particular focus on university repositories that contain a range of different digital resources. The remaining two terms functionality and user group will now be examined. Functionality will be studied with particular emphasis on the value placed on different types of electronic resources in relation to the objectives of the repository. This will include a review of the literature on content studies in IRs. For user group, the focus will be on usage and a review of the literature on usage of electronic resources.

Functionality

IRs are created with different objectives and purposes in mind and therefore, their *functionality* is different. The functionality refers to the main motivation behind the creation of the repository and consequently its main purpose or function (Heery and Anderson 2005). Examples of repository functionality are: increased access to discovery of electronic resources, preservation of resources, new modes of dissemination and/or publication, institutional asset management and promoting sharing and re-use of resources.

There are four ways of understanding the role or functionality of IRs (Chan 2004):

1) As one of the best ways to provide access to the results of scholarly funded research in order to maximize its impact.
2) As a way to increase and diversify the digital scholarly materials that are available for research, teaching and learning.
3) As an efficient means of increasing the visibility of an institution's digital output and serving as a type of showcase.

4) As necessary and essential infrastructure for reforming scholarly communication and publishing.

Relationship between functionality and content

The actual content type of an IR will vary according to its perceived functionality. Although there are a number of drivers behind the creation of IRs, and there can actually be several at the same time, there are, as we have seen, two main views on institutional repositories (Lynch and Lippincott 2005). The first regards IRs as an alternative path to Open Access publishing (Crow 2002; Ware 2004a) and essential to reform the current scholarly publishing system (Crow 2002; Hubbard 2003). In this view, institutional repositories are considered an important tool to enable the transformation of scholarly publishing towards Open Access. The content type of these repositories is generally related to formal publishing, in particular journal articles. They tend to focus on e-prints, whether they be pre or postprints. This outlook on IRs tends to concentrate on changes in the form of *access* to scholarly materials rather than the nature or type of material that is available as this is e-print orientated. Repositories in this sense are mainly outside the scope of this research as we are focusing particularly on non-formal e-resources. This is however, currently an actively researched area as the debate about the future of formal scholarly publishing and the impact of Open Access continues.

The second trend defines IRs as essential infrastructure for scholarly communication. IRs are a way of acknowledging the importance of these resources and providing a means for organizing, disseminating and preserving them. In this view a university-based IR is a "set of services that a university offers to the members of its community for the management and dissemination of digital materials created by institutions and its community members". (Lynch 2003:328)

These types of IRs tend to contain a much broader range of materials, including but not limited to e-prints. These IRs broaden the scope to include any type of scholarly digital resource produced by members of the institution. Each IR in turn will limit or refine this scope according

to their particular intentions. The objectives of repositories however may be multiple and often contradictory. Some for example, may have as a primary aim Open Access and the collection of e-prints but will be willing to accept other types of content. There may be other repositories that collect a large array of materials, including e-prints, but do not intend to destabilize the scholarly publishing system, or may not even have Open Access on their agenda at all.

To date, however, there are few systematic and focused studies on the types of contents within IRs, despite its importance. It is quite likely that this is due to the fact that most repositories are fairly recent and most of them have relatively little content (Jones, Andrew et al. 2006; McDowell 2007). It is not certain yet however, if reaching a critical mass is just a question of time or an inherent problem with IRs. There are only a few studies on the types of contents within IRs and these are reviewed in the following paragraphs.

Content and item types in IRs

Prior to creating an IR at the University of Edinburgh, a study was conducted to discover what types of electronic resources were being put online by academics on departmental and personal websites. (Andrew 2003) Although a substantial amount of material was available online, it was widely dispersed over different sites and therefore not easily found. A systematic approach to resource discovery was taken by revising each webpage in turn starting from the main page and following links to other pages. Originally the intention was to only document formal research material but this was readjusted to reflect the scope of the material found which tended to be broader, in particular for the Humanities and Social Sciences. Examples of types of resources found are sheet music, public lectures, data sets, newsletters, maps, Scottish witchcraft database and sound clip archives. An unexpected find was the relatively low number of preprints on personal web pages, from all subject areas. The author suggests that this could be due to the success of eprints repositories in some subject areas, such as arXiv. Another possible explanation is that the majority of articles are linked to from the academic's online CV.

Andrew suggests that academics may prefer to link to the final refereed copy as the final published article is more appropriate for a CV.

The MIDESS Project carried out between 2005 and 2007 in order to explore the use of institutional repositories for the management of multimedia content also reported a large number of multimedia resources (what in this study are being referred to as non-formal electronic resources) such as learning objects, medical slides, digitized film clips, coin collections and digitized medieval manuscripts. Again they found that although these types of materials are quite widespread within an institution they are scattered among different platforms ranging from virtual learning environments to academic home or course pages (Emly 2007).

As part of a broader study on IR development, Ware analysed a total of 45 institutional repositories in detail (around 42,700 documents). In terms of content distribution Ware found that 22% of items were eprints (both pre and post), 20% theses and dissertations, whilst 58% were categorized as others, and included mainly grey literature (reports and working papers) as well as a large collection of digital images. One of the conclusions is that the type of data "varies considerably, but from inspection it appears that copies of final published articles make up a relatively small proportion*"* (Ware 2004a:118). Ware proposes that there is no evidence to suggest that IRs are reforming scholarly publishing. He concedes that most IRs are still in the early stages of development, although even the ones that have been around for longer have only collected a small fraction of formal research output. This study shows though that more than half the content is what we have called non-formal e-resources.

However, of these non-formal e-resources most of them appear to be textual based, rather than more complex digital objects. "Documents are mainly text-based articles of various types, there is currently little evidence of more complex digital materials, datasets, etc. although some IRs state that they are planning to allow these at a later stage" (Ware 2004b:25). Ware suggests that IRs could actually act as a tool to collect the scattered and unaccounted for digital objects that

are produced by academics and simply placed online. By placing them in IRs these resources would be easier to discover and use. "Digital objects on university networks are currently largely uncatalogued, widely scattered and not managed. There is no central catalogue or database for such materials. Such materials, are therefore difficult to discover and use, and difficult to keep track of and preserve. Mitigating such collections to a central IR would address some of the issues" (Ware 2004b:20).

In 2005 the first broad and systematic survey of IR deployment was conducted (Westrienen van and Lynch 2005). Thirteen nations were surveyed about IRs in their country. By means of a questionnaire they were asked to estimate the number of IRs in their country, the average number of documents, software deployed, disciplinary coverage and other questions related to national policies on IRs . They were also asked to indicate the relative percentage of documents by content. The list of content types they were asked to estimate were: Articles; Books and Theses; Primary data; Video, music,etc.; Course material; and Other (namely). There is no indication how this list was decided.

Norway, Sweden and Belgium, reported a larger percentage of theses and books and three countries, France, Italy and the UK reported a prevalence of articles. The Netherlands reported 40% theses and 20% articles, but no further percentages are given. Australia reported a higher percentage of non-formal material (83%), namely primary data. US data was reported in a separate article, and although there are no percentages they point out that the study found a "significant number of institutions are committed to institutional repositories that go far beyond e-prints" (Lynch and Lippincott 2005: unpaginated). Additionally this study lengthened the list of content types adding for example: newspapers, data sets, digitized institutional assets from library and museum collections, exhibitions, performances, interview transcripts, maps, plans/blueprints, software and laboratory protocol. There is no data for the remaining four countries.

It is highly likely that the figures for the previous studies are not particularly accurate, as noted in the article itself (Westrienen van and Lynch 2005). A major problem was the definition and use of terms related to IRs which affected answers on the questionnaires, depending on the interpretation of certain terms. An example of this is the number of items held within a repository as what is meant by a record differs. They report that in the US, it is assumed that a record will contain the digital object, whilst for example in the Netherlands, a record can refer just to the metadata. This naturally makes a big difference in terms of counting items. The Netherlands reports that the average number of records per IR is 12,500 but when refined to include only records with the digital object the number is 3,000. Although the survey provided initial interesting points, the authors note that further work is required.

A key finding of the survey is the difference in the main drivers behind the creation of the IRs between countries and the impact this has on content collection. Except for the US and Australia, they found a strong emphasis on textual material for the IRs with the focus still being very much on traditional publications. In terms of the role that IRs play or could play in the future of a more advanced networked scenario of digital data one of their main points is that: "We did not hear issues raised about the need to manage, preserve and provide access to large, complex, inherently digital objects such as datasets, software, simulations and the like that constituted fundamentally new forms of scholarly communication not accommodated by the existing scholarly publishing system. We did not hear about the impact of e-science and e-research on scholarly communication" (Westrienen van and Lynch 2005:unpaginated).

In particular in the US, IRs are seen as a much more general-purpose infrastructure, especially when looking towards the development of e-science. In addition, they note that a large number of IRs are built with DSpace which they argue is better for managing diverse resources in comparison to EPrints which is better for e-prints and more popular for example, in the UK (Probets and Jenkins 2006). Interestingly Ware, finds a similar US-Europe divide with regard to IRs and content type (Ware 2004b). It would be interesting to see if this is a trend or a

circumstance that will change as IRs are further developed as well as comparing this situation with the development of IRs in other countries.

These findings however, do not necessarily lead to the conclusion that European institutions are not interested in non-formal electronic resources. European nations, in particular the UK, have strong centralized subject repositories for data, such as the AHDS (Arts and Humanities Data Service) for the Arts and the Humanities and the Resource Discovery Network. This type of infrastructure is not as strong in the US, and this could be the cause for a different approach to content type of IRs (Lynch 2003; Westrienen van and Lynch 2005). However, since the publication of these surveys the UK Arts and Humanities Research Council withdrew funding for the AHDS arguing that many of the objectives, such as building information communication technology (ICT) expertise, had been achieved. Additionally they argue that the "long term storage of digital materials and sustainability is best dealt with by an active engagement with HEIs rather than through a centralised service" (AHRC 2007). It would seem that they are advocating for a more institutional type repository approach to Arts and Humanities resource management. But how concerned are IRs with long term preservation and, more importantly, how well designed and concerned are they for adopting materials that are not eprints or ebooks? It also begs the question about the relationship between IRs, data repositories and subject-based archives. It will be increasingly important to address these issues if non-formal e-resources are to be properly managed, disseminated and preserved.

In 2007 there was a follow up to the 2005 US survey on IRs (Lynch and Lippincott 2005) done by McDowell (McDowell 2007) on evaluating IR deployment in American academe. One main difference was that in contrast to the 2005 survey, the term institutional repository was explicitly defined. Repositories should be institutional (therefore departmental repositories were not considered) and should collect and provide access to diverse faculty output. Another main methodological difference was that rather than sampling through survey replies, the researcher specifically selected the repositories that fulfilled the selection criterion. These were found by

consulting repository listings. As the current repository landscape develops, tools for finding and listing repositories are becoming important and necessary for research. There are currently several repository listings, with different degrees of coverage and criteria for selection. The Methodology section of this thesis will analyse different directories and access their strengths and usefulness.

As with the previous US study, this research looked at the proportion of the types of materials in repositories. The contents of the IRs tracked where categorized under the following content types: ETDs (electronic theses and dissertations); e-prints (pre or postprint articles); working papers and technical reports; conference proceeding and presentations; e-journals and e-books; learning objects; multimedia files (digital audio/video); datasets; pictures (images); digitized archival documents and university records (historical texts and primary resources); non-scholarly institutional publications; undergraduate student work; graduate student work (non-ETD); and course content (syllabi, assignments, lectures). The study recognizes that the resulting figures are approximations only and that both the categorization and the counting require fine-tuning.

This survey found that over 40% of the content was student produced work in the form of electronic theses and dissertations. Both formal and informal scholarly output accounts for 37% of the content. Of this about 13% are e-prints (pre and post) and e-books, a little over 20% of grey textual literature in the form of working papers and technical reports and only about 1% was more informal type resources such as conference presentations, learning objects, podcasts and datasets. Scanned images were reported separately and were responsible for about 13% of all items in the repositories. Finally administrative materials (categorized as non-scholarly materials) such as newsletters, guides, agendas, minutes make up about 4.5%, followed by 3% for historical textual documents (usually digitized archival material).

The list of categories for different material types was based on the original Lynch and Lippincott survey and also on another IR survey conducted by the ARL (Association for Research Libraries) , although no further details are given. The ARL survey conducted in 2006, and published as Spec Kit 292 (Bailey, Coombs et al. 2006), surveyed 87 member libraries. Of the respondents 43% has an operational IR, 35% were planning one and 22% had no IR or plans to develop one in the immediate future. This survey also found that the most common type of materials were ETDs followed closely by articles. They also found a large percentage of conference presentations, technical reports and working papers. They argue that it is not surprising that there is such a widespread inclusion of grey literature as this type of material is relatively easy to include in IRs and in general these do not have already robust publishing avenues. In a sense they are the 'low hanging fruit'.

In 2007 the MIRACLE Project (Making Institutional Repositories A Collaborative Learning Environment) carried out a census of IRs in the US (Rieh, Markey et al. 2007) focusing on five key issues, one of them being IR content. This survey differs from others in that they categorized responses according to different stages of repository development. Respondents were asked to estimate the percentage of materials within their IRs according to "three dozen digital types" but no further detail is given. The results show that for fully implemented repositories the most common types of materials were PhD theses, working papers, postprints and raw data files. Repositories in the pilot stage also listed theses but mentioned preprints and learning objects as other common types. The study found that both groups rarely gave high estimates for more complex and non-textual materials such as e-portfolios, software, sound recordings, interviews transcripts, maps, etc. This data appears to contradict the other previous US survey (Lynch and Lippincott 2005) but in both cases figure are too approximate and the categorization of resources too vague to be able to draw any definite conclusions. Both studies serve as ground breaking initial approximations for IRs in the United States.

A recent SPARC survey (Barkier 2008) aimed at gathering information from repository managers about their predictions for IR development, in particular for trends in content collection. Respondents indicated that they thought it likely that there would be more incorporation of student research, non-academic content (examples are alumni material, planning and development documents, enrolment data) and community based content (examples are commencement addresses, lectures and papers, material developed in collaboration with non-affiliated members such as regional agencies and NGO's).

These surveys reviewed tend to only touch upon the subject of repository content, among other points, and none are solely focused on content type, but rather offer an overview of several aspects of IRs. There are no systematic studies of institutional repository content, and further work must be done to identify, specify and map the repository landscape (Heery and Anderson 2005). Precisely because we do not understand what is contained within institutional repositories it is even more difficult to understand what IRs are being used for. This point is crucial, as identifying IR use is critical if we are to understand their role in scholarly communication and publishing. It would also aid further development of IRs, to ensure that they are useful.

Target user group and usage

Heeney and Anderson's fourth typology is *target user group,* referring to the intended users of the IRs. Potential users of IRs are learners, teachers and researchers (Heery and Anderson 2005). Some IRs are more targeted towards particular types of users, whilst others are more broad and include users outside the academic setting.

In the case of IRs for academic users, it is important to differentiate between two types of users of IRs: academics as creators of resources and academics as readers or users of electronic resources. Most scholars will belong to both types, but their motivations, priorities and needs

are very different. These two 'natures', academics as authors and academics as readers, may lead to conflicting interests or contradictory behaviour in their attitudes towards using content and making content available. We shall refer to the first types as academic authors or creators (depositors) and the second type as academic readers (users).

Academics as creators (depositors)

In the formal electronic publishing arena there has been discussion regarding author motivation for self-archiving and/or publishing in Open Access journals (Nicholas, Huntington et al. 2005a; Salo 2008). Although it would seem that there is a wide support for Open Access, almost all eprint repositories have found a real problem with participation (Ashworth, Mackie et al. 2004; CENLFEP Committee 2006; Jones, Andrew et al. 2006) as scholars as authors do not tend to voluntarily self-archive. This has been so marked, that it has even led certain Open Access advocates to insist on self-archiving mandates in order to make the movement possible (Harnard 2001; Harnard, Carr et al. 2003; Pinfield 2005; Sale 2007)[18]. One study found that institutions with self-archiving mandates will significantly increase the number of articles deposited (Sale 2006). However, it can be argued that institutional repositories will only be successful if the community adopts and uses them voluntarily and not because they are obligated. On the other hand, other studies have found that 95% of researchers would archive if required to do so (Sale 2007).

Davis and Connolly (Davis and Connolly 2007) interviewed faculty members as part of a study on the reasons for non use of Cornell's institutional repository. They found that the service is largely under-populated and unused by the faculty. Some of the main reasons mentioned by academics for not depositing in the IR were: redundancy with other modes of disseminating information, the learning curve, confusion with copyright, fear of plagiarism, inconsistent quality, and concerns about whether posting a manuscript constitutes "publishing". Many

[18] For a complete list of institutions or organizations with self-archiving mandates see: http://www.eprints.org/openaccess/policysignup/

academics were already making their work available either through their web page or a disciplinary repository and did not see the need to use the institutional repository. The study concluded that the crisis in scholarly publishing, acutely perceived by the Library community, is regarded as a non-issue for most members of staff. This may help to explain the apparently inexplicable attitude of academics to self-archive their articles despite some evidence to suggest that it increases the visibility of their research. It is quite likely that many academics are satisfied with the *status quo*.

In another study on depositors (Thomas and McDonald 2007), participation patterns in repositories were measured and compared by looking at how many items were deposited by author in a particular repository. They found that author participation as a depositor is generally wide spread but shallow. Repositories tended to have a large number of authors that deposited only one item. There are however, a number of limitations to this study. The focus was only on e-print type repositories that used EPrints software. Of the initial candidate set of 838 repositories, only nine met the necessary criterion for analysis. Additionally the study assumed that the depositor was necessarily the author of the paper. There is little evidence to suggest that self-archiving is popular and anecdotal experience seems to imply that a great deal of material in institutional repositories has been deposited by mediated self-archiving (usually done by the Library staff). This strongly contradicts the assumption that depositors will be the authors of the item.

There appears to be no further work done yet regarding the behaviour and motivations for authors to create and deposit electronic resources in institutional repositories or even for posting their work online in general. A study found on self-posting research material noted this gap (Andrew 2003). Understanding author creation and deposit or posting behaviour is key factor to building up a better picture of author use of IRs.

Academics as readers (users)

A main driver for institutional repository development has been to make academic research output available to a much larger community by eliminating access barriers and thereby increasing its visibility and impact. As we have seen repositories offer services to store, organize and maintain the institution's digital research output. In addition repositories aid online discovery of digital materials by assigning standardized metadata to items and supporting the OAI-PMH, thereby facilitating resource discovery by search engines and users. The question is are resources within institutional repositories actually used?

Determining the use of electronic resources

Usage studies in the print world are generally based on re-shelving data, questionnaires or citation analysis (Jamali, Nicholas et al. 2005). Online digital resources have created new opportunities and challenges for measuring use. As in the paper world, to date most online usage studies have focused on more formal publications, in particular electronic journals (Nicholas, Huntington et al. 2005b; Rowlands, Nicholas et al. 2007). This is not surprising considering publishers and libraries are particularly interested in usage studies of electronic journals due to both their scholarly and economic importance (Jamali, Nicholas et al. 2005). In the past few years there has also been a growing interest in measuring use of electronic books through surveys and log analysis (Nicholas, Huntington et al. 2007; Rowlands, Nicholas et al. 2007; Nicholas, Rowlands et al. 2008).

Citation analysis

One approach to understanding the use of digital publications has been to apply traditional bibliometric techniques, such as citation analysis, to the online world. This method in particular has been applied to usage of articles in e-print repositories. Early studies indicated increased impact of a paper when made available freely online (Lawrence 2001), have been followed by other studies looking at citation impacts of preprint archives (Harnard and Brody 2004). These studies suggest that the free availability of preprints increases the frequency at which articles are cited. However, once again the focus has been on formal publications (journal articles), rather

than the repository materials as a whole. Citation analysis is a method designed for formal publications and it would be difficult to apply to informal textual documents as these do not usually have the same rigorous and regular citation formats as formal publishing and even more so to the current heterogeneous world of non-formal e-resources.

Log analysis

Another method for measuring use is by analyzing the logs that servers use to record and store details about the user's interaction with the system. Known as transaction log analysis (TLA) this methodology has been used for over a decade to measure the use of online digital resources (Jamali, Nicholas et al. 2005), including electronic journal usage (Jamali, Nicholas et al. 2005; Nicholas, Huntington et al. 2005b; Huntington, Nicholas et al. 2006; Bollen and Van de Sompel 2008b). These studies have focused on the use of formal electronic publications usually contained within the publisher's publishing platforms. There is still a relatively small amount of studies that use TLA for measuring scholarly journal usage (Jamali, Nicholas et al. 2005) but its popularity is increasing (Nicholas, Huntington et al. 2005b). Log analysis has also been used for ebook usage studies (Nicholas, Huntington et al. 2007).

The study on the reasons for the non-use of Cornell University's IR focused mainly on interviews with eleven faculty members for data collection but also used log analysis as a small part of the methodology (Davis and Connolly 2007). By looking at the server logs they found six outlier resources that were particularly popular. What is particularly interesting to note is that all these resources are what could be considered non-formal resources. Two resources are video (one the biography of a Noble prize winner in Physics and the other a lab demonstrations which is used on a course). Two resources are scanned images of class books that are no longer in print. In both cases these are used in college classes and there are several links to them from other organizations and society websites. The remaining two highly popular resources are reports from librarians. One report is a particularly controversial one on Open Access and the other on the changing nature of the library catalogue. This log analysis provides interesting

insight into the nature of the most popular resources in the repository, but is treated marginally in the paper and no further analysis is offered.

Another case study from the IR at the University of Nebraska-Lincoln also discovered through analysis of download logs that some of the most popular resources were "a number of works that had been or logically would be deemed unsuitable for ordinary (i.e.) paper publication" (Royster 2007). Notable examples are a collection of early American texts, mainly out of print and some that were never even published at all. The researcher in charge of collecting, editing and publishing these in the IR dedicates a great deal of time to offering reliable editions of these texts and this work is regarded as valid scholarly output. Another popular resource is a dictionary of invertebrate zoology that although accepted, prepared and peer reviewed for publication was never published as the publisher deemed it too costly due to its size and limited commercial appeal. And finally the recreation of the musical program played during a five-month exposition in 1898 that allows users to explore the different pieces of music and musicians over this period. Royster argues that IRs can act as primary publishers of these types of resources providing they also concentrate on outreach or publicity of the material as their popularity indicates perceived usefulness by users.

A refinement on the TLA methods has been the development of deep log analysis, which combines log analysis with more qualitative methods such as interviews and demographic information. Deep log analysis has been used mainly for research in electronic journals, e-books, health information systems and digital libraries (Jamali, Nicholas et al. 2005; Nicholas, Huntington et al. 2005b; Huntington, Nicholas et al. 2006; Nicholas, Huntington et al. 2006). This method is described in more detail in the Methodology section.

The LAIRAH (Log Analysis of Internet Resources in the Arts and the Humanities) project recently used deep log analysis to evaluate the use of electronic resources in the Arts and the Humanities, focusing on subject-based repositories (Warwick, Terras et al. 2006; Warwick,

Terras et al. 2006). They studied three services: AHDS, Humbul and Artifact[19]. The AHDS acted as a repository: storing, preserving and providing access to digital resources in the Humanities. The 15-month long study used the server transaction logs to identify well-used and neglected digital resources. An aim of the project was to discover what factors influenced the use (or non-use) of digital resources. The study found that between 30-35% of resources were not used, resources in popular subjects were extremely well used and some key resources were used intensively by a small, specialized community.

A key finding of the study was that it was surprisingly difficult to gather the logs from the different servers (Warwick, Terras et al. 2008). Logs are often undervalued, not maintained or made available. In the case of one service they could not provide the logs because they did not have the necessary technical expertise. Another service, that had integrated personalization features was concerned about data protection and worried that users could be identifiable. In order to provide the logs, they stripped them of additional data for analysis and this caused unforeseen time delays in the study. These types of problems have not been reported from other log analysis studies of scholarly resources. However, as we have noted these have generally being done on logs from publisher's servers who will generally have integrated information systems and ample technical expertise.

The problem of collecting usage data from disparate sources is currently being addressed by MESUR (MEtrics for Scholarly Usage of Resources). This is a two-year project to investigate metrics derived from the network-based usage of scholarly information. They are specifically looking into how to aggregate usage data across multiple scholarly information resources. These hybrid-metrics include publishers but also take into account grey literature and other objects that exist outside the realm of scholarly journal publishing (Bollen and Van de Sompel 2008a).

[19] AHDS (Arts and Humanities Data Service) http://www.ahds.ac.uk/. The funding for this service from was discontinued in April 2008. Humbul and Artifact were merged to become part of Intute http://www.intute.ac.uk/artsandhumanities/

In the case of IRs, there do not yet appear to be any log analysis usage studies. This methodology could contribute to our understanding of IR use, which is currently not very deep. Indeed both IRs and subject based repositories must be studied from a user perspective (Heery and Anderson 2005) and understanding the use of IRs is the next crucial step in the development of this field.

Visibility

The use of a resource on the online world can be affected by its visibility. If something is particularly difficult to find, then its likelihood for being used (and useful) is reduced. The visibility of electronic resources is currently difficult to gauge. How visible are items within institutional repositories to, for example, search engines? Commercial search engines attempt to crawl and index as much of the publicly indexable web (PIW) as possible. This PIW comprises all web pages that can be crawled by search engines and that do not require registration or authorization. For academic resources, for example, materials placed under password protected websites or subscription services, such as electronic journals, are usually not indexed by search engines. The exception is Google Scholar that appears to have a number of deals with major publishing house to index their full text content (Vine 2006). Another significant portion of the web that is not properly indexed is what is known as the deep or hidden web. These are web pages that are created dynamically usually generated by queries to content databases and are therefore not indexed effectively by search engines.

Two studies were found that look at the visibility of the metadata harvested from OAI server providers (i.e. mainly institutional repository items) to search engines. The first study (McCown, Liu et al. 2005) harvested 776 OAI repositories for a total of over 9 million records. The referring URL of the items were extracted from the Dublin Core field URL. In other words, the URL of the item itself, not the URL of the metadata page. They then queried three main search engines in order to see if these URLs had been indexed. The results show that Yahoo!

had indexed 65% of the URLs, Google 44% and MSN 7%. They attribute the higher coverage percentage by Yahoo! to an agreement between the search engine and a harvester provider called OAIster[20]. They do not however know if this agreement still exists or whether this is how Yahoo! indexed the repositories in the first place. Detailed information about commercial search engine indexing strategies is difficult to obtain.

This study was followed up in 2008 (Hagedorn and Santelli 2008) although in this case they only conducted the research for Google. They found that 44.35% of the harvested URLs were indexed by Google, which shows no improvement on the previous study figures. Additionally they mention that in April 2008 Google withdrew its support for OAI-PMH for its sitemaps, which suggests that this situation will not improve in the foreseeable future. Site maps allow websites to inform search engines about the URLs that are available for crawling and aid them in the discovery of hidden or deep URLs.

So although repositories and OAI-PMH were set up to aid users in the discovery of mainly academic digital resources, we still have little understanding of how effective this has actually been. The studies found present rather disheartening results but they do show that agreements with search engines or possible modifications to IRs software setup and sitemaps could dramatically increase the results. This is an important issue as stated by the British Academy: "In current discussions, the questions of access visibility for readers are too often ignored. This would not matter is the repository was treated only as a piece of backroom mechanism, designed to help authors preserve their materials while leaving the author to deal with the questions of how his materials are found. But many discussions seem to be based on the view that the repository is itself the primary access point to its content" (Spark Jones, Bennett et al. 2005:82).

[20] OAIster is reviewed in more detail in the Methodology section.

In a study on disciplinary coverage in IR, Zuber found that one of the first obstacles to resource visibility was actually being able to find the repository on the university's website. He notes that "the lack of visibility within an institution's own website speaks to poor recruitment and incentive strategies" (Zuber, 2008:unpaginated). A CIBER report on new journal publishing models also found low levels of awareness of IRs among academics (Rowlands and Nicholas 2005). Repository managers have focused keenly on advocacy for content collection but if one of the main objectives is to make resources more visible and accessible, then they must begin to face outwards as well.

The LAIRAH report found difficulties in discerning whether a resource is non-used because it is not perceived as useful or because the potential users cannot find it or are not aware of its existence (Warwick, Terras et al. 2006). They conducted a workshop on perceived usefulness of digital resources. Using data from the log analysis, a combination of well-used and neglected resources were selected and presented indistinctly to participants. Participants were then asked to determine whether they thought a resource was popular or neglected (Warwick, Terras et al. 2008). The study found that participants were quite critical of resources and tended to assume that these were neglected, even those that were well used. It is quite likely that users of digital resources have become accustomed to digital resources provided by commercial publishers, large libraries, archives and museums and see this as the expected standard. Resources produced by academic specialists may be content rich but lack professional interface and technical design. Users seemed to favour digital resources produced by institutions or organizations that they knew and trusted in the offline world. An example is the Imperial War Museum Concise Art Collection, created by the Imperial War Museum.

In another study (Warwick, Galina et al. 2008) producers of popular electronic resources were interviewed in order to try and discover common characteristics for successful digital projects. The study found that well-used digital resources were usually produced in institutions that supported and encouraged digital projects. Producers of the resources were provided with an

adequate level of technical support, had a research assistant with a good level of subject and technical expertise and most importantly, they all worked hard on promoting the digital resource. These dissemination activities were varied but included sending emails and flyers and presenting at conferences, workshops and seminars. The study notes that "interviewees stressed that both producers of digital resources and funding agencies must realize the key role of dissemination for a project's success" (Warwick, Galina et al. 2008:389). It seems to be that when materials are produced outside the formal channels of communication, producers most make up for a lack of a formal publisher by actively engaging in the promotion of the resource. Are institutional repositories a useful and additional channel for resource distribution? Can IRs aid in increasing the visibility and the dissemination of these resources?

Deep log analysis can indicate the provenance of users, but more information would be useful for determining whether resources can be discovered by potentially interested parties. An important perceived function of publishing or making a resource available online is for increasing visibility and impact. Some research (Lawrence 2001; Harnard, Brody et al. 2004; Eysenbach 2006) has found that Open Access or freely available articles in repositories, are more highly cited (and supposedly therefore more visible). The visibility and impact of articles is measured by citation rates. The same criteria are not as applicable to electronic resources as these are not cited in the same way as journal articles. So how can the visibility of non-formal e-resources be measured?

Stemming from Bibliometrics, the field of Webometrics or Cybermetrics studies "the quantitative aspects of the construction and use of information resources, structures and technologies on the www, drawing on bibliometric and informetric approaches" (Björneborn and Ingwersen 2004:1217). A novel approach has been to attempt to understand the digital environment and the impact and use of digital materials through link analysis. Link analysis is a methodological approach for looking at web-related phenomena. The rationale is that a link to a particular resource can reveal important information about its usage or perceived importance.

In particular, website interlinking is related to a range of informal types of scholarly communication (Wilkinson, Harries et al. 2003), which is one of the main interests of this research. Linking to a resource acts as a reference, similar to citations in print publications (Björneborn and Ingwersen 2004; Thelwall, Vaughan et al. 2005), or at least serves as an indication of usage or perceived usefulness. Moreover, further studies have shown hyper linking motivation to be possibly more complex than citations (Kim 2000; Wilkinson, Harries et al. 2003) and potentially a novel source for understanding informal scholarly communication (Wilkinson, Harries et al. 2003).

Link analysis has been used as a research method for a variety of studies in particular for the scholarly information environment. Studies have been done on: research productivity and impact and link counts (Thelwall 2003b; Thelwall and Harries 2003; Thelwall and Harries 2004), research collaboration and academic relationships (Payne and Thelwall 2004; Stuart, Thelwall et al. 2007), invocation of scholars on the web (Cronin, Snyder et al. 1998) and digital libraries (Zuccala, Thelwall et al. 2007). It has also been used for the wider web in a study on social networking sites (Thelwall 2008).

In the repository landscape the previously mentioned paper the IR at the University of Nebraska-Lincoln on the popularity of non-formal resources (Royster 2007), indicates that online links towards the resources has positively affected downloads. Important traffic drivers are bibliography and reference listings from credible scholarly organizations. Examples given are Penn University, the MLA international bibliography and Intute.

Link analysis is also used to produce one of the indicators for the beta version of the Ranking Web of World Repositories[21]. The aim of the ranking is to measure the global visibility and impact of repositories and is developed using a combination of web indicators. The four

[21] Ranking Web of World Repositories. Produced by Cybermetrics Lab, Consejo Superior de Investigaciones Científicas (CSIC), Spain. For more information see: http://repositories.webometrics.info/about_rank.html, May 2009

indicators are: size (the number of pages extracted), rich files (the number of .pdf files extracted), scholar (the mean of the normalized total number of papers and recent papers published) and visibility. Visibility is calculated using link analysis and refers to: *the total number of unique external links received (inlinks) by a site*. The visibility indicator weighs 50% of the ranking system. The ranking focuses on repositories with formal publications and deliberately excludes repositories that are devoted to non-scientific papers or archival material.

There appear to be no studies done so far specifically on institutional repositories using log or link analysis. The few studies done so far indicate that there is plenty of scope for research and that measuring usage and assessing the visibility of electronics resources is a promising area.

Institutional repository definition

Using the Heerey and Anderson typology (Heery and Anderson 2005), this study therefore concentrates on IRs as defined by a broad and inclusive *content type*, an institutional *coverage*, in which the *functionality* is set of services for managing and disseminating electronic resources and in which the *user group* is defined as scholars, but divided into scholars as authors and scholars as readers.

In addition, IRs will share generally speaking, the following characteristics:

- web-based repository of scholarly material
- cumulative and perpetual (a record)
- open and interoperable, using OAI complaint software

and thus be part of the scholarly communication process by collecting, storing and disseminating (Ware 2004a).

Although there is general notion that repositories somehow contribute and possibly even alter scholarly communication and publishing by facilitating new modes of publication and peer review as well as increasing data sharing by enhancing access to resources as well as the re-use

of raw data and learning objects (Heery and Anderson 2005), there is little understanding of how this actually works. IRs are still relatively new and it is probably only a matter of time for a significant corpus of literature on the subject to build up. In the meantime however, the large body of literature on the effects of electronic publishing, both of formal materials as well as electronic resources in general, on scholarly communication and publishing, are a useful starting point for understanding the possible implication that IRs may have. The following section will review the key issues as these are most likely relevant to the impact of IRs.

Electronic publishing and its implications for scholarly communication and publishing

The study of scholarly communication and publishing

The pursuit of academic knowledge is defined as both a cooperative and cumulative activity; researchers will generally read others work and base their own research on previous results. In turn, they will seek to disseminate their findings to the interested community and thus produce material for the further accumulation of knowledge. Over time, the academic community has relied on a series of different methods for communicating both in written form - letters, newsletter, journals, emails - as well as orally - meetings, conferences, seminars, talks, telephone, video conferencing and others. This has developed into a highly complex international structure currently in place today. As scholarly communication has expanded and increased, it has also became more complex. Over the past decades, in order to constantly improve information provision and overall communication, it has become increasingly important to study the scholarly communication structure in a systematic way (Vickery 2000).

Formal and informal scholarly communication and publishing

Traditionally, scholarly communication has been divided into two channels: formal and informal (Meadows 1998). The division of formal and informal communication has served as a

general guideline for studying and analysing scholarly communication. In addition, formal scholarly publishing has been strongly associated with the journal and book formats. However, the introduction of networked computers and electronic publishing, has altered and blurred the boundaries between these traditional divisions and formats. This has important implications for scholarly publishing and of course for electronic publishing and its analysis, "basic assumptions about categories of information will need to change. For example, the distinction between formal and informal communication channels sits uneasily with the use of computers and networks" (Meadows 1998:37). Some studies tend to tacitly assume that these channels are unaltered.

Attitudes towards electronic publishing

The subject of electronic publishing has been treated with great skepticism, pointing out the difficulties with accessing quality, permanence, copyright, costs of computers and networks, among others; and with great enthusiasm, prophesizing glorious changes and a brilliant future with cheaper, faster, democratic, universal and better forms of publishing.

Early literature on electronic publishing in particular, tended to be quite controversial. The medium was criticized, mainly because the nature of the material is not reliable (it is easier than print for anyone to publish), it is not fixed (there is no authoritative and definitive version), it is volatile and not permanent (URLs changed, publications disappear from one day to the next), it is uncomfortable to read on screen, dubious quality (i.e. not peer reviewed) and there is no easy way to annotate the text (Grenquist 1997). Other work (Fillmore 1993; Odlyzko 1995; Ginsparg 1996; Peters 1996; Adair 1997; Grenquist 1997; Unsworth 1997; Wheary and Schutz 1997) concentrated on the more positive aspects and described the technology as a medium which would greatly enhance communication and publishing, in particular, as it offered numerous possibilities that are not available in print format.

Some of the common concerns have diminished over time and are now rarely a factor (such as slow internet connections, lack of internet connection, unreliable networks, among others). However, it should be noted that in many parts of the world these are still important factors and by no means is electronic publishing a global or universal reality. The relative importance of other concerns has certainly diminished. It seems that certain technological aspects (such as reading on screen or annotating text) will be resolved either with improved technology in the future or with the effective combined used of print and electronic technologies. Reading on screen is slower and digital text is harder to comprehend (Ramirez 2003; Cameron 2005) and long articles are general printed out for deeper reading and annotating. However, users favour the availability of texts in digital format (Ramirez 2003) and reading preferences will vary according to the objectives (Liu 2005). Some electronic publications that do not adapt well to the printed medium are generally read on screen with ease. Examples are more interactive sections of online newspapers, blogs, wikis, and others. As online publishing has become more stable URLs have tended to stay the same, and other initiatives such as the DOI (Paskin 2003), have attempted to mitigate some of these problems, although with still limited success.

The biggest unsolved issues, which are still being researched and discussed today are more related to the quality of the material that is published, the role of publishers and libraries in scholarly publishing, new electronic forms and their impact on scholarly communication, preservation and continuity of electronic forms, copyright and intellectual property, to name a few. These issues are also key to electronic resources and institutional repositories.

Is it published?

Some of the earliest literature on electronic publishing is on the status of the material, which was online and available. Initially electronic versions were not considered worthy or serious publications just because of the nature of their distribution format. Publishing is generally defined as making something public. However, Internet technology enabled an individual for

the first time to distribute a document widely with relative ease. So, what does publishing actually mean in a networked environment?

The technology initially caused confrontation between parties regarding the concept of publishing. A close examination of editorial policies for four journals or scholarly societies found great differences regarding the status of articles posted on the Internet; ranging from considering them as published previously and therefore unacceptable for submission to regarding the Internet as simply a form of distribution more akin to posting copies by mail than a serious form of publication (Kling and McKim 1999).

A similar situation was found with the Virginia Tech's online Electronic Theses and Dissertations (ETDs) distribution system, with many supervisors warning their students against the system as their thesis might be considered 'published'. The author recalls phoning publishers in order to ascertain their position but many of them did not yet have a policy in place (McMillan 1999).

As publishing on the Internet became more popular among academics and the information systems used for searching and retrieving these articles became more prominent and available the situation has become more confusing, to such an extent that in 2002 the UK's JISC (Joint Information Systems Committee) launched the RoMEO (Rights MEtadata for Open archiving[22]) project in order to investigate the rights issues with regard to posting articles on the web and self-archiving in repositories (Gadd, Oppenheim et al. 2003a; Gadd, Oppenheim et al. 2003b; Gadd, Oppenheim et al. 2003c).

This situation is still unresolved. As seen previously, eprint repositories generally speaking, are created in order to affect the scholarly publishing system. In the case of IRs, that contain diverse types of electronic resources, it will be important to understand their role, especially in relation

[22] For more information see RoMEO website (http://www.lboro.ac.uk/departments/ls/disresearch/romeo/)

to publishing. Will electronic resources within IRs be considered as published? It is clear that the range of electronic resources in IRs is very diverse and they have been created for very different purposes. So, what characteristics must an electronic resource have in order to be considered as published? In order to understand this, we must first understand the role of publishing and what it means within the academic community.

What is publishing? A few proposals

An important step towards understanding the nature of electronic publishing is by discerning and defining the characteristics of a publication (Boyce 1999; Kling and McKim 1999; Frankel, Elliot et al. 2000; Kling and McKim 2000; Tenopir and King 2001; CENLFEP Committee 2005). The role of the publisher in publishing is key but what role or roles does a publisher play? Why and what are these functions important for? Can this role be undertaken by the authors or their institutions? What is a publisher in an electronic environment? The following proposals concentrate both on the role of the publishers (their function) and what publishing means.

An International Working Group (Frankel, Elliot et al. 2000) put forth a proposal that defines the desirable characteristics of an electronic document in order for it to be considered an electronic publication in Science. The report argues for permanence, persistence in integrity of appearance and completeness of content, public availability, version control, authenticity, notification, persistence of location, author's commitment not to withdraw, quality control and archiving and long-term preservation. The report focuses on electronic online journal publication and regards the final published article after peer review as the crucial fixed point- a Definitive Publication.

According to Boyce (Boyce 1999) print publishers disseminate information as widely as possible, provide a system for accessing information, ensure clarity and effectiveness of

presentation, create a system of standardized exposition and produce material in long-lasting format. When considering electronic publications Boyce discusses the role, not of the publisher or the article, but of the paper journal and the numerous purposes it serves including- providing information about and for the community, keeping abreast of latest results, representing the corpus of a discipline's knowledge (in some cases), certifying an author's credentials, serving as a record of progress and setting the standards and defining scholarly norms of what makes an acceptable research paper. This definition is strongly linked to journal publishing in particular.

Tenopir and King (2001) argue that publishers provide three basic services: they collect manuscripts according to the interests of a particular readership; ensure quality and they provide distribution and access. Together with libraries they also offer indexing, abstracting and bundling articles together in categories. In addition, publishers also provide marketing and advertising services to promote the dissemination of the publication (Warwick 2002).

The above definitions however, are somehow limited because they tend to concentrate on particular formats for publication, such as journal articles and continue to understand publication in print-based terms (Kircz 2001; Kircz 2002). In particular they do not help towards understanding new digital forms that are online and are currently not defined.

A more useful definition understands publishing as more than just communication. Kling and McKim make a key distinction between publishing as a communicative practice in which the main priority is to be read by your intended audience and publishing from a functionalist perspective, where it serves to allocate status, allocate resources and communicate results (Kling and McKim 1999).

Publishing as a communicative practice

As a communicative practice publishing is used by authors so that their work may be widely read and credited. Within the academic community there are differences in the disciplinary practices. A socioinformatic approach (Kling 1999; Kling, Rosenbaum et al. 2005) emphasizes the differences in disciplinary practices and criticizes the work that treats the academic community as homogeneous. They argue for field-specific valuation of different document formats (i.e. journals, reports, conference proceedings) and the publishing venues for them as well as their perceived value.

Research on electronic publishing tends to treat the academic community as homogeneous with similar communication patterns throughout. What may be true in electronic publishing for one academic community, such as the use of preprints, is not necessarily a norm in other disciplines. For example, the case of the eprints server arXiv in Physics as a successful model for self-archiving (Harnard 2001) has been criticized because it disregards the fact that unlike Physics, not all disciplines will have a history of preprints as a form of communication (Kling and McKim 2000). Disciplinary perspective is very important, both for the producer and the user of electronic publications. It is highly likely that this will apply to electronic resources, as different subject areas, agree on different communicative forms and channels. In addition, the uptake and use of IRs will also be affected by disciplinary differences.

In their model of scholarly publishing as a communicative practice electronic publishing is seen as a continuum and there are different degrees of publication. This view is more flexible and accommodates different types of digital materials at different stages within the publication process. The degree to which a document is a published can be measured by its publicity, trustworthiness and accessibility (Kling and McKim 1999; Kling and McKim 2000). Accessibility refers to the ease with which a document can be located and obtained. IRs have partially been set up to address this problem by offering stewardship and long-term preservation of academic resources. Publicity is the degree to which interested readers are aware of the

availability of a document. Kling and McKim argue that "publicity does not automatically and inexorably proceed from a document's availability on a global network" (Kling and McKim 1999:905). IRs and OAI-PMH can make resources more available to search engines and other retrieval mechanisms but how important is their contribution? Will IRs have to engage in other types of activities, more akin to publishers, in order to promote the use of their resources? Trustworthiness refers to the different signs that scholars use to determine the relative value of a document. Peer review is one indicator but there are others and these can vary between disciplines and in particular, for more informal publications users may rely on other indicators such as the producing source or author name. Will IRs, as part of a university, increase the trustworthiness of a non-formal e-resources?

As publishers and libraries have entered the electronic publishing world, their defined roles have also become slightly blurred. Unsworth refers to *pubraries* and *librishers*, in which they take up tasks belonging to the other. For example, pubraries, such as ProQuest, will offer their publications directly to the end user, whilst librishers, such as the University of Michigan Scholarly Publishing Office offer electronic publications. The dangers of course, is that they do not take up *all* the tasks, and for example, pubraries usually lack attention to issues such as preservation and aiding users, whilst librishers are still learning marketing, distribution and working with author skills (Unsworth 2005).

Publishing from a functionalist perspective

From a functionalist perspective (Kling and McKim 2000), electronic publishing has important implications on the economics of publishing, the allocation of status (peer review and reward system) and the role of scholarly publishing in the communication process in general. These three general areas overlap greatly and changes in one have effects on others. Publishing cannot be seen in isolation from the role it has in academic life and any change will have an impact on the general system (Day 1997).

From the onset the mere possibility of electronic publishing as opposed to traditional print publishing challenged and questioned the role of each one of the players in scholarly communication and publishing- authors, editors, typesetters, printers, distribution agencies, subscriptions agents, librarians, teachers, researchers and students (Willis 1996). Many of the ensuing discussions have been a response to changing degrees of influence or control. It is clear that electronic publishing has somehow forced members of the academic community to take a deep look and evaluate their role within the university and its purpose. Why do we publish at all? What function is being met? What changes when we publish information in a different form or medium than before? So far, in this new digital environment, these are questions that are still unanswered. In the case of formal electronic publishing, one of the most important debates has been centred on the functions and economics of publishing, with important topics such as the serial crisis (Tuttle 1989; King and Tenopir 1999; Houghton 2001), Open Access (Harnard 2001; Chan, Cuplinskas et al. 2002; Brown, Cabell et al. 2003; Bullinger, Einhäupl et al. 2003; Harnard and Brody 2004; Suber 2004) and alternative business models (Willinsky 2003; Harnard, Brody et al. 2004; Holmström 2004) widely discussed. In the case of electronic publishing in general, including electronic resources, the role of publishing for allocating status, ensuing quality via peer review and the general reward system are important functionalist aspects that must be discussed.

Rethinking the role of publishing in academia – quality, status and reward

Electronic publishing has not only affected the economics of publishing, it has also affected the perceived role of publishing within the scholarly community. Within scholarly communication, publishing is seen as a way for researchers to communicate their results to interested colleagues. Researchers publish primarily to communicate information: "for the advancement of knowledge, with attendant benefits to their careers and professional reputations" (Ginsparg 1996:unpaginated). However, others argue against this storybook version of science (Helmut 2000) that regards scientists as selfless individuals who communicate and publish only in the

pursuit and advancement of knowledge and science as a purely *cognitive connection*. If we understand publishing within a social system, within a reward and prestige framework and not only for communicating, then the role of electronic publishing become more complex. Kling and others (Kling and McKim 1999; Kling 2000; Kling, Rosenbaum et al. 2005) argue for a Social Informatics approach to understanding information technologies not just as information tools but in terms of their associated structures and politics. Radical changes in technology do not necessarily imply changes to customs, patterns of behaviour or the social structures of science and scientific communication.

In his article 'Why Do We Write Stuff That Even Our Colleagues Don't Want to Read' (Humphreys 1997) points out that the current reward system, which places the publication of monographs as a requirement for tenure as well as being an important status symbol, contributes to the current information overflow. The sciences follow a similar trend with the 'Publish or Perish' syndrome. Both lead to a deluge of information- on one hand, books which could have too little to say – and on the other hand, scientists that squeeze out as many articles as possible from a single research experiment. As reward systems require more publications from the academic community, competition becomes fiercer, publications become highly specialized and library budgets are stretched. As pointed out by Thatcher (2005:unpaginated): "The conflict between library practices and tenure committee requirements is one more instance of the failure of universities to examine the logic of their institutional systems—puzzling in view of the university's self-image as the bastion of rationality."

In the journal publishing world the importance placed on the Journal Impact Factor for the evaluation of researcher's work has come under scrutiny (Seglen 1997; Hecht, Hecht et al. 1998). There appears to be a growing discontent with the current form of evaluation. The existence of a large corpus of electronic journals has led to the proposal of new or complementary forms of evaluation, especially based on hyperlinking and visibility or web

impact factors (Kim 2000; Wilkinson, Harries et al. 2003; Thelwall and Harries 2004; Aguillo, Branadino et al. 2005).

With regards to the scholarly monograph the Modern Language Association, with wider applications for all the Humanities, has made a call for universities to review their tenure requirements as currently there is a strong bias towards requiring a monograph publication (Jaschik 2005). The panel proposal recognizes that there have been important changes in publishing, which are partly due to technological innovations. It is increasingly harder for a young scholar to publish a monograph and this leaves many potentially good candidates without a chance. Publishers are immersed within a market and their interest in the diffusion of knowledge is based on making profits. On occasions high quality work coincides with a demand for it and in others, especially with the scholarly monograph, it does not. As a panel member reports: tenure opportunities "should never depend on the vagaries of the scholarly publishing market" (Jaschik 2005:unpaginated). This action can be seen as a stance against a reward system, which is based on commercial and economic activities rather than the content itself. The report argues the need to "explicitly change their expectations such that there are 'multiple pathways' to demonstrating research excellence" (Jaschik 2005:unpaginated).

For academics that decide to innovate with new electronic forms of doing or communicating research the situation is complicated further. In a study to find out what made digital resource projects successful Warwick (Warwick, Galina et al. 2008) found a clear correlation between institutional context and creation and use of digital resources. In institutions where digital humanities research was recognized and valuable and the creators gained prestige and promotion from their work, it was much more likely for other similar electronic resource projects to be developed. Interviews with creators of successful and well-used digital resources mentioned the importance of institutional backing. It is highly likely that where the creation of digital resources is considered outside the main research activities of staff, there will be less work in this area.

The main issues of 'certification' and 'validation' are as yet unresolved. Publishers have always played an important role in this area providing legitimacy to an author's work. As well as challenging publisher's role as disseminators of information, electronic publishing is also challenging their role as gatekeepers and certifiers of knowledge. For most publications this boils down to peer-review and its importance within the system. Journal publishing relies on peer-review as a measure of assured quality. Peer review has also been amply debated and it is clear that there are pitfalls to the system, especially with a large number of authors competing to get into top impact factor journals. It is however, regarded frequently as the only way of certifying a publication. Former editor of BMJ notes "Despite a lack of evidence that peer review works, mot scientists (by nature a sceptical lot) appear to believe in peer review. It's something that held 'absolutely sacred' in a field where people rarely accept anything with 'blind faith'" (McCook 2006:27).

A plausible hypothesis for understanding this apparent dogma is that peer review is more important nowadays for evaluation purposes of the author rather than that of the publication. "In many fields, the principal use of peer-reviewed journals is not to publish research but to provide apparently impartial criteria for universities to use in promoting faculty" (Arms 2002:unpaginated). Third party institutions evaluate researchers work through the publishers. In this sense, where you publish and how often becomes the metric through which rewards are determined. Peer-review may not be about certifying a publication, or at the very least not just about certification, but it is a part of the much wider system that has been discussed previously. A change in peer-review implies changes for the whole hierarchical system.

With electronic publishing some authors (Odlyzko 1995; Harnard 2001; Odlyzko 2002; Henry 2003) have argued for new forms of peer review and certification although there is no consensus in the actual form. Other research reports a unwillingness from the scholarly community to abandon the current system (Swan and Brown 2003). The form or forms that

peer-review can take in the electronic publishing world is probably one of the thorniest issues in the revised literature. However, assuring quality and validity in publication is a matter that can be considered unresolved both in the print and electronic world with numerous examples over the years of misinterpretation and outright falsification of data. It seems best to avoid sweeping generalizations about the goodness and effectiveness of all peer review.

In addition, there are different grades of perceived quality. "Scholars do not treat all peer-reviewed reports as equally trustworthy; rather they rely upon a variety of processes and markers, which are dependent upon everything from the structure of the discipline itself to the social networks that the readers are embedded in" (Kling and McKim 1999:905). In addition the type or form of a publication is also important for determining perceived quality and this varies between disciplines. "Scholarly communities have developed conventions about the relative status of different paper documents. This is reflected in highly differentiated category systems of books, journals, reports, conference proceedings, working papers, and field specific valuations of these documentary formats (and publishing venues) (Kling and McKim 1999:996). As new online documents appear, that have no print paper-based equivalent, such as blogs, wikis, datasets, scholarly hypertextual editions, virtual reality models, among many others, the academic community will most likely continue to expand and develop their ideas on publishing, in order to cope with these new formats. The role of IRs as a means to collect, preserve, certify and provide access to this growing number of resources, will be essential to understanding their role within the scholarly publishing and communication systems.

Institutional repositories and scholarly communication and publishing

As we have seen there are two main background reasons for setting up IRS. The first is to attempt to modify the current scholarly publishing system, and tend to support the Open Access movement. In these cases IRs are a strategy to improve access to traditional scholarly content (Andrew 2003), in particular eprints. The SPARC position paper proposes a disaggregated

scholarly publishing model divided into four components: registration, certification, awareness and archiving. They argue that IRs can play a more active role in these processes and thereby breaking the publisher's monopoly. Currently publishers are responsible for registering, certifying and together with libraries provide awareness and archiving functions. The paper argues that IRs can register, certify, provide awareness and archiving for eprints. These functions are summarized in Table 2, with the processes, actors and process sponsors shows.

Function	Process	Actors	Process Sponsor
Registration	Posting electronic paper to repository	Academic author- researcher	Repository sponsor
Certification	Peer review Associative certification Online response	Academic referees Academic referees Academic respondents	Overlay journals Academic departments Repository sponsor
Awareness	Interoperable open repositories and support services	Librarians	Academic institutions Professional services Third-part providers
Archiving	Perpetual access	Librarians	Academic institution

Table 2 - Scholarly communication functions in a disaggregated model

Lynch however, makes an important distinction between scholarly communication and scholarly publishing, which he notes are generally used interchangeably in IR8 discussions (Lynch 2003). Publishing is a subset of scholarly communication which is broader, more diverse and sometimes more informal. He believes that IRs do not give universities a new publishing role but that they do offer a new way of dissemination of scholarly communication. Publishing is of course much more than just dissemination but it is also very limited in the *genres of communication* that it can handle (mainly articles and books). Therefore IRs, that handle a wide variety of material, will not affect scholarly publishing as much as they will improve scholarly communication. The effects of IRs are threefold: they empower faculty for the dissemination of their digital materials, motivate preprint dissemination (in particular where

there is no subject repository) and finally they will "encourage the exploration and adoption of new forms of scholarly communication that exploit the digital medium in fundamental ways" (Lynch 2003:332).

Lynch regards IRs as instrumental for new types of scholarly communication. He states that IRs "can support new practices of scholarship that emphasize data as an integral part of record and discourse of scholarship. They can structure and make effective otherwise diffuse efforts to capture and disseminate learning and teaching materials, symposia and performances, and related documentation of the intellectual life of universities" (Lynch 2003:332). The creation of IRs as institutionally backed services will guarantee preservation and access, which are a prerequisite for scholarly legitimacy. This in turn will clear the ground for discussions, within each discipline, about the relative value of these new forms of communication, in particular for evaluation, tenure and promotion.

Lynch's ideas about using IRs to validate new forms of scholarly communication is a useful one, although he does not extend this idea to new forms of publications. However, if we consider that IRs can serve as a type of test bed for new forms of scholarly communication, it might then be possible for these forms to become new forms of publications, as they become more ubiquitous and established. IRs can serve as an intermediate step for new types of publications, until they enter the formal realm by being recognized and acknowledged as such, providing they are useful. Possible each discipline may develop new genres of publishing that satisfy particular communication needs.

Print on paper has served the academic community remarkably well over the centuries as a fixed and permanent format for publishing. However it is important to remember that these formats have not always been the same and have developed over time in order to meet different needs. "As a channel for conveying research information, printed books and journals have therefore changed appreciably with time. The way they look now depends on the nature of the

research and the history of the research community. It, therefore, embraces a mix of factors ranging from the expectations of the community to developments in the printing technology". (Meadows 1998:119). It is relatively difficult to determine what types of publications may develop, or how current publications may evolve. There is some research that looks into new ways of communicating and publishing research. The future of scholarly publishing and communication in a networked environment is examined with closer detail in the next section.

Towards new forms of electronic publishing

Future publishing forms

Some literature has made attempts at describing what the next step in electronic publishing will be and what we can expect of technology and people. It is worth emphasizing that our perceptions of the correct way of presenting information and writing and communicating research results are learnt and relate more to a sociological aspect than a scientific one. Our way of writing is still based on the perceived notions of good writing from the print world. The slowest, most painful, changes will be to understand the digital publishing world as allowing us to communicate research in a way that has never been done before and therefore hard to imagine. Work has been done that reveals the inadequacies of the current system and the potential in electronic publishing to salvage them, calling for further development beyond the digitalization of print paradigms (Odlyzko 2002; Henry 2003; Van de Sompel, Payette et al. 2004; Warner 2005). This trend focuses on using technology in order to create innovative ways of doing, communicating and distributing research.

Nelson and Maly see the final article as an abstract of a much wider body of work, which they define as a pyramid of scientific and technical information (STI). The print paradigm requires this synthesis due to space and cost constraints of the medium. Although this form is useful for certain purposes they note that the research done to produce this article left a trail of information that can be useful for future analysis. In the print world this material is usually lost to all but a

few colleagues: either filed away in a cabinet, stored on a hard drive or simply forgotten or thrown away. The remaining material is artificially segregated and all interrelationships are lost (Nelson and Maly 1999). They see Digital Libraries as simply reproducing print paradigms and separating STI into different media formats when this is technologically no longer necessary. Technology should allow for the trail of information and its relationships to remain.[23]

Van de Sompel et. al. argue that our concept of a unit of communication can no longer be limited to the article format and they propose to revise it in both a technological and a systems sense (Van de Sompel, Payette et al. 2004). Electronic publications are not simply print clones with hyperlinks and a bit of multimedia inserted (Warner 2005). An electronic publication can be in itself a dataset, a video, a text, a hyperlink; as long as in itself it is sufficiently comprehensive to convey meaning. Modularity should be a model for electronic documents (Kircz 2002). Released in 2008, the OAI-ORE (Open Archives Initiative – Object Reuse and Excahnge) specifications were designed to address the issue of describing and exchanging aggregations of web resources[24]. These compound digital objects, as they are also known, combine different distributed electronic resources with different media types (text, images, data, video). An example, of this would be an audio of an interview that also includes the textual transcription (a Word and an XML version) and a series of photographs taken of the interviewer and interviewee. ORE was created to address these issues of managing relationships between different objects.

In the monograph world, scholarly monographs should not be thought of as books. According to some the book-like shape of the monograph is there to meet publishers requirements and not scholarly needs. "In many ways they never were books (…) If you look inside the average scholarly monograph, it is almost entirely support structure, a very bony fish: review of

[23] One interesting example of this has been the work done by the Internet Archaeology Journal in conjunction with the ADS at York, which stores relevant data from the journal's articles in the Archaeology Data Service, ensuring access and long-term preservation to this electronic resource.

[24] For more information see the Open Archives Initiative – Object Reuse and Exchange website (http://www.openarchives.org/ore/)

literature, chapters on methodology, bibliographies, appendices, extensive footnotes. The flesh may be sweet but there is precious little of it" (Arnold 1993:unpaginated). He also argues for a modular approach using electronic systems. The fact that the scholarly monograph has not gone electronic is not a failure. There is no point in reviving the monograph or seeking to keep it alive in the electronic environment rather this is the opportunity for scholars to assess the current scholarly information process and to re imagine it into the system we want. However, in recent years there has been a lot of work on ebooks (Ramaiah 2005) and although currently there is there is little, rather outdated and contradictory literature on the current situation of scholarly monographs (Armstrong and Lonsdale 2000), this situation seems set to change.

The overall argument is for a reduction in the unit of communication and an expansion in the relationships this unit can have with other units. In the print world, the physical nature of the object forced us to artificially create units (articles, journals) and citations and bibliographic references attempted to keep truncated relationships alive. Electronic publishing does not have these physical restrictions and there is no need to import them into the digital world. It is necessary to reformulate our ideas with regards to scientific communication, and how we present and formulate our ideas and arguments into something which is more effective (Kircz 2002). In this sense we are blurring the lines between what was traditionally regarded as a formal or informal publication as well as the boundaries of formal and informal communication.

Blurring the boundaries- the future of informal publishing in an electronic environment

It is important to note that the dividing line between formal and informal publishing has never been clear-cut, even in the print world. Once again, there are important disciplinary differences. Scholars place different values on forms of publication depending on their disciplinary background. As pointed out by Kling, the nature of the publications and the expectations of a publication vary between fields. "For example, while both computer science and biology rely upon conferences extensively, computer scientists value conferences as a final publishing

forum. In contrast, biologist typically do not, viewing them as a more informal forum for sharing results. Many humanities disciplines, such as literature or history, value books as a publication forum, while the lab sciences typically devaluate book and book chapter publication. In many areas of physics, talks hold a high formal status whilst most disciplines use talks primarily for informal communication" (Kling and McKim 2000:896).

In this sense the distinction between formal and informal publishing has never been clear-cut as the relative value and therefore its degree of acceptability as a valid publication format varies between disciplines. However, with the advent of electronic publishing, this distinction has become more blurred and in some cases is actually almost illogical. Electronic publishing has further blurred the boundaries between the formal and informal (Ramalho and Castro Neto 2002) and in our distinction between data, information and knowledge (Meadows 1998).

The main interest of this thesis is to discover the implications of repositories on the visibility and use of electronic resources outside the framework of formal electronic publishing. In traditional terms these would generally be referred to as informal publishing although this term is no longer particularly useful as many of the main characteristics of informal publishing in the print world are more difficult to adapt to the electronic world. "Discussion about the value of electronic documents is often hampered by the fact that it starts from what is usual in the paper world and attempts to impose that on an electronic environment" (Kircz 2001:266).

There is plenty of scope for research in this area. The academic community feeds from and into the informal chain of information dissemination resulting in numerous types of materials and publication, that although useful, are rarely named or studied and even less in the electronic world. Houghton et al. point out that: "One of the features of the literature on research information and communication practices is its uneven emphasis, with some areas (e.g. the use of journals) covered extensively and other (e.g. informal dissemination of material) rarely examined in detail. The same holds true with articles and e-journals" (Houghton, Steele et al.

2004:231). The main focus has been on scientific journal, especially with regards to cost models (Henry 2003; Houghton, Steele et al. 2004; Van de Sompel, Payette et al. 2004), less on monographs and practically no attention on non-text material, born digital objects or informal mechanisms of dissemination. And yet the wider implications of electronic publications are most likely affecting what has traditionally been known as the informal realm of scholarly communication and publishing.

Within the research done regarding electronic publishing and its implications for scholarly communication it has been important to analyze the role that publishing, both formal and informal, plays within the system and how this is modified by an electronic presence. Several authors have worked on examining the role of scholarly publishing. One of the most important things to note is that technological changes have come far more rapidly than sociological ones that usually take longer.

This should not stop us though from examining alternate proposals for electronic publishing, in the form of institutional repositories and the electronic resources contained within. It is important to understand the impact that these can have on scholarly publishing and communication in general. As with any future-telling activity, the literature aptly describes the current changes needed but has more difficulties in assessing and conveying possible solutions and there are still many unresolved issues. In this sense there is an open field for research into these issues to help develop, guide, construct and evaluate the necessary framework for the future.

Summary

This chapter provides an overview of the literature for three main fields: electronic resources, institutional repositories and scholarly communication and publishing. Different approaches to definitions of electronic resources are reviewed, in particular non-formal e-resources. This is

followed by a brief overview of the origins of institutional repositories. Different drivers behind the creation of IRs are examined through the relevant literature. Following the repository typology approach issues related to the perceived functionality of an IR and its content makeup are addressed, focusing on the sparse content studies to date. We then look at usage studies, in particular at two methodologies, log analysis and link analysis. Studies that employ these methods for measuring use or the visibility of electronic resources are reviewed. Finally we look at IRs in the wider context of scholarly communication and publishing and the role that e-resources may play by looking at the literature on the role of electronic publishing and its future.

The review shows that there are important contradictions in the perceived role of IRs for scholarly communication and publishing. It seems that the drivers for creation are directly linked to the content but there is little understanding of the types of materials that are actually deposited and stored in IRs, especially for non-formal e-resources. There is plenty of scope for further research in this area in order to understand what role non-formal resources and institutional repositories are playing or could play in the future for scholarly communication and publishing.

The next chapter will describe the Methodology used to address the research questions.

Chapter 3- METHODOLOGY

The purpose of this chapter is to describe the approach to the research design and to discuss the different issues in relation to the methods and procedures employed for data collection and analysis and the relationship between the various datasets. The chapter explains the reasons for a more qualitative rather than quantitative approach to the research design in particular within the broader context of Social Informatics. The different issues regarding the research methodology, in particular the use of surveys, interviews and link analysis as appropriate methods for data collection are discussed in conjunction with the approaches towards data analysis including sampling, grounded theory and link analysis. A critical overview of current repository listings as tools for data collection is provided.

Introduction

As discussed in the literature review, scholarly communication and publishing, and to a lesser degree, informal scholarly publishing and communication are developed areas of research within the LIS field. There currently exists a solid body of literature on various aspects of this academic endeavour and activity. The aim of this research is to study non-formal electronic resources within the institutional repository context and suggest possible implications for scholarly communication and publishing. The literature review showed that non-formal electronic resources and their use within institutional repositories is a relatively unexplored area. The main focus of previous research has been on formal electronic resources rather than on non-formal. This parallels the print world where it has been shown that there is more research interest in formal publishing than in the so-called grey literature. This is not surprising given the predominant importance of formal publications in scholarly communication. However, as we have argued, with the advent of the networked environment the role of non-formal electronic resources is changing, in particular with initiatives such as e-Science that place more emphasis on the importance of datasets and other types of research outputs. This confirms the need for more research to be conducted so that we can better understand exactly

how these non-formal electronic resources will function within the scholarly communication and publishing system.

Institutional repositories are also a relatively unexplored research topic, although not necessarily to the same degree as non-formal electronic resources. One of the main reasons is the relative newness of IRs and the fact that as information systems these are still in development. As we have seen in the literature review the proposed objectives and perceived benefits (such as increased dissemination and access) or disadvantages (such as reliability of content) of IRs are still under discussion and a great deal of IR literature deals with these issues (Johnson 2002; Pinfield 2003; Ware 2004a; Ware 2004b; Kircz 2005; Awre and Baldwin 2006; Flores Cuesta and Sánchez Tarragó 2007). As the number of IRs grows worldwide and agreements and consensus are reached, it is more likely that a larger number of empirical systematic studies will be conducted. To this date most studies have discussed issues surrounding the characteristics and potential roles of IRs (Crow 2002; Lynch 2003; Chan 2004; Heery and Anderson 2005; Awre and Baldwin 2006) or have tended to be more descriptive case studies describing their set-up (Ashworth, Mackie et al. 2004; van der Kuil and Feijen 2004; Davila, Núñez et al. 2006; Muller, Ulrich et al 2008). With the number of IRs increasing globally, it follows that wider, more comprehensive and empirical studies need to be conducted. The 2005 seminal census studies on IR deployment in 13 nations (Lynch and Lippincott 2005; Westrienen van and Lynch 2005), with a 2007 follow up for the USA (McDowell 2007) are an indication of this. However, there is still much work to be done in the field of empirical studies on the current state of IRs.

Consequently, this particular study is timely and comes at a moment when it can be argued that a sufficient amount of IRs currently exits to merit deeper insight into the repository landscape. What are currently lacking are evaluative frameworks for studying IRs that can be adapted and used even within a changing digital environment. Research on IRs is currently moving away from prescriptive literature on what the effects of IRs will be towards more empirical based work on what effects IRs are actually having.

This mix of palpable research gaps leads to a relatively unchartered repository landscape that can be ideally studied with an exploratory approach to research design (Williamson 2002). This research does not aim to question or prove developed theories but to offer an initial insight into aspects of IRs that have not been studied before. The following section describes the approach taken in the research design and the different issues involved.

Research design

In the LIS field researchers have successfully employed quantitative and qualitative methods to investigate the various aspects of both the print and the digital information environment (Williamson 2002; Beck and Manuel 2004; Gorman and Clayton 2005). The so called *paradigm wars* between supporters of quantitative only or qualitative only approaches to research are relatively exhausted (Bouma 2000:175). Both approaches are now viewed as useful depending on the issues to be researched without the need to engage in the often circular debates that dominated social sciences methodological discussions during the 60's and 70's. Moreover, it is increasingly common to see studies in which researchers employ methodological resources from both approaches and combine them at different phases of the research process (Tashakkori and Teddlie 1998; Punch 2005). As stated by Williamson these two methods can be "complementary and, in combination, give both the broader, larger-scale picture, as well as a more detailed understanding of a specific situation" (Williamson 2002:p.35)

During the development of my own research I found Punch's notion (2005) that quantitative and qualitative aspects of research may be viewed as applied on a continuum particularly useful. Punch's simplified model of research is divided into pre-empirical and empirical stages (see Figure 2). The two different research approaches -qualitative or quantitative- will define both how linear and structured the application of this model is, as well as the types of data collected. Quantitative research is generally considered to work with pre-specified research questions,

tightly structured design and pre-structured data whilst qualitative research uses general guiding questions, loosely structured design and no structure is pre-imposed on the data (Punch 2005:23). Qualitative and quantitative definitions are applied on a continuum and the research approach can lie anywhere on a spectrum between these two extremes.

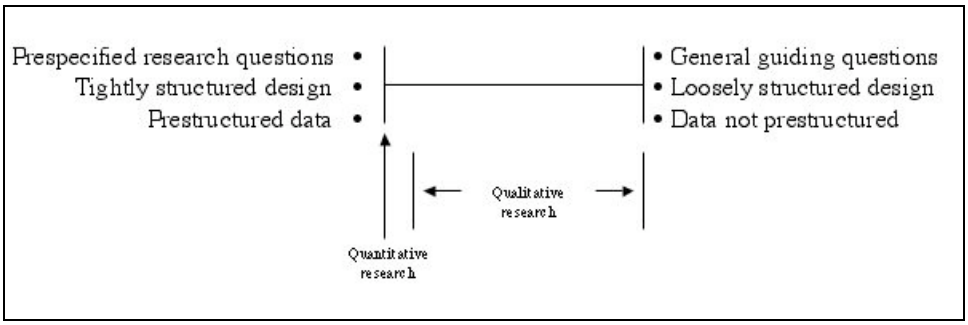


Figure 2 - Simplified model of research (Punch 2005)

Initial background desk research carried out for the development of this research proposal had already suggested that non-formal electronic resources and IRs were relatively unexplored areas. The exploratory nature of the research called for a broad look at the phenomenon being investigated (Bouma 2000:91). A tightly structured approach would artificially limit the research scope and it was important not to impose pre-existing expectations on the research. Therefore, a more qualitative flexible research design was considered more appropriate. This approach to the research process is generally associated with an interpretivist rather than a positivist approach.

Positivism is understood here as the attempt to "apply scientific methods to the social sciences, and is most usually associated with deductive reasoning and quantitative data collection" (Williamson 2002:37). Deductive reasoning develops arguments from general instances to particular ones, usually through a hypothesis testing approach to research. This type of method is strongly associated with a scientific or positivist approach to social science research that claims that social sciences can be investigated in the same way as natural sciences, and that all

scientific knowledge is based only on what can be objectively observed and experienced (Williamson 2002:27).

Interpretivism, on the other hand, is more concerned with "meaning". The social world differs from the natural world in the sense that our social world is interpreted or constructed by people and such construction is expressed as a production of "meaning" that is not universal, but historically and socially determined. "Interpretivist researchers regard their research task as coming to understand how the various participants in a social setting construct the world around them" (Williamson 2002:31). This approach is more closely associated with qualitative approaches to research although quantitative data may also have a role.

The focus of this research is on the implications of a particular type of information system for the academic community and its activities. It involves understanding the relationships between different actors, such as repository managers, depositors, library staff, university administrators, publishers, authors and readers, within the academic context and in relation to the deployment of institutional repositories. One of the key issues from the literature review on scholarly publishing and communication was the importance of the sociological aspects of these activities and to this end this research is concerned with the opinions and interpretations of the participants. The sociological dimension in information studies has been previously acknowledged in the literature (Wilson 1981; Kling 2000; Robbin and Day 2006; Cronin 2008) and as stated by Cronin: "The socio-cultural dimensions of knowledge and the socially embedded nature of information and communication technologies (ICTs) are, and to some extent always have been, integral to the theory base of information science, an assertion that is easily confirmed by inspection of the published literature" (Cronin 2008:467). The sociological approach towards information studies differs from another popular approach that emphasizes the technological over the social.

Social and technological approach to Information studies

Research in the information studies field, especially that concerned with the implementation and deployment of information systems, may tend to focus more on the performance or impact of the system itself rather than the sociological aspects surrounding its implementation. The extreme case is when information systems are viewed solely as a conglomerate of equipment, data flows and procedures; in short the emphasis is on their role as information tools. Harvey (2005) notes that in this context, the most obvious uses of research in information environments are for:

- Problem solving
- Development, evaluation and improvements of services and systems
- Provision of information before introducing new systems or services

However, another approach to information studies is to highlight the sociological aspects of information systems implementation and performance. The literature review showed that many of the main issues surrounding IRs are more sociological than technical (Awre and Baldwin 2006; Davis and Connolly 2007). Understanding the implementation of information communication technologies (ICTs) through social aspects has been used in different fields such as information systems, information science, computer science, sociology, political science, education and communications (Kling, Rosenbaum et al. 2005). The main problem is that these studies stem from different fields of enquiry and have generally been published in different disciplinary journals (Kling 2000; Kling, Rosenbaum et al. 2005; Robbin and Day 2006). This makes it particularly difficult to track down all the relevant literature and to identify a cohesive body of literature on which to build further research. In the late 1990s a workshop at Indiana University addressed this issue and suggested a common umbrella term, Socio Informatics (SI), to group this type of research and which they defined as "the new working name for the interdisciplinary study of the design, uses and consequences of ICTs in ways that take into account their interaction with institutional and cultural contexts" (Kling 2000:218). Socio

informatics is defined more by the problem area that it addresses than by similar approaches in theories or in methods.

Socio informatics argues that often, tacit assumptions are made about the potential benefits of ICTs that do not take into consideration the more complex and subtle social phenomena. The danger of viewing ICTs as information tools that lead to or cause specific social changes when implemented is that they tend to streamline processes that are in practice usually much more complex. A socio informatics approach is an alternative means to direct effect theories and considers social and technological aspects as interrelated. Social changes usually take much longer, and are deeper and more complex than accounted for by technological determinism. Socio informatics recognizes that the implementation of information systems can have both positive and negative effects on organizations and in many cases these are unexpected.

In the case of the academic environment, we can think of the organizational structure of universities as a highly structured and hierarchical network. As shown in the literature review, although IRs are implemented as technological tools their development has important social consequences, especially in the area of publishing and communication. Some of the literature, in particular that of Open Access, is explicitly political. However, IRs may also have unexpected impacts that have yet to be studied. A socio informatic approach towards understanding who deposits and with what types of materials, could shed light on the more profound effects of IRs. For example, self-archivist evangelists have been at a loss to explain the academic community's reluctance to self-archive despite what they describe as obvious advantages. A socio informatic approach could offer useful explanations. "These power orientated explanations differ from those explanations that simply focus on the advantages an ICT can offer to some groups" (Kling, Rosenbaum et al. 2005:31). The research design, data collection and analysis in this thesis will draw on some of the key socio informatics issues discussed here.

Research process

This study, therefore, employed a mainly qualitative approach. The next step was to incorporate the data collection and data analysis methods into the research design. It was decided that the best approach was to use a combination of data collection methods and employ these different datasets to construct the current institutional repository landscape in terms of non-formal electronic resources. The data for this study would be collected using survey and case studies that would include interviews and usage data, in the form of link data. An initial overview of the general research process is presented followed by a detailed look at each research method.

The literature review was used to identify key concepts for the study and to inform the reiterative design of the research questions. Although initial research questions and objectives were established prior to undertaking the literature review, it was assumed that an important part of the research process would involve reworking and fine-tuning the research questions as the main issues and concepts were gradually revealed through the collection and analysis of data. The first stage of the data collection in the form of an online survey for repository managers, was to give a general overview of the repository landscape and aid in the further identification of key issues providing initial insight into content typology, workflows, repository objectives and repository usage. This was followed by seven repository case studies. Each case study included an in depth interview with the repository manager to dig deeper into the main issues identified from the survey data. Furthermore, link data were collected from all case study repositories to find out how the resources within the repository were being used. Figure 3 is an overview of the research design.

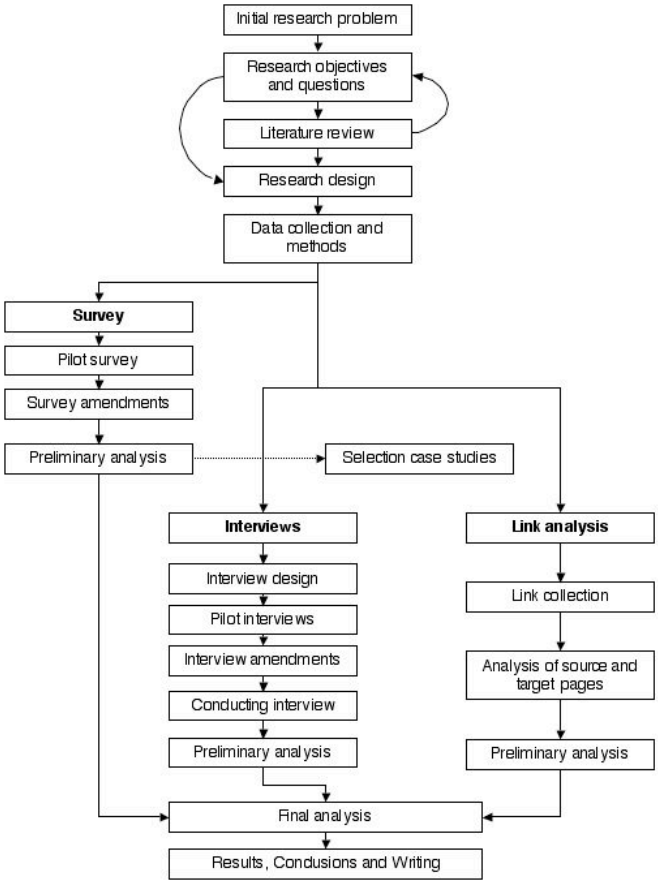


Figure 3- Overview of the research design

The research was carried out over an eighteen-month period as shown in Table 3.

Literature Review	September 2006 – September 2007
Online survey	Data collection - July to September 2007
Case studies	Repository manager interviews – February-March 2008
	Link analysis – March 2008

Table 3- Research methods and timeline

Each stage of the methodology is described in greater detail in the following sections.

Research methods used

As mentioned previously the research methods chosen were: a literature review to reiterate research questions, data collection using an online questionnaire for repository managers, followed by seven case studies collecting further data from repository managers using interviews and usage data in the form of links.

Literature review

The literature review may be used as an integral part of the research process in particular when the research topic is relatively young, providing an initial foundation for a new research topic (Levy and Ellis 2006). As mentioned by Hart "Analysing the literature can have as much intellectual and practical value as collecting first/hand data. A thorough critical evaluation of existing research ideas often leads to new insights by synthesizing previously unconnected ideas, and can provide methods for the collection of data and suggest solutions tried in similar situations" (Hart 2001:2). In the particular case of this research, the systematic overview of the literature helped further refine the key topics and also flagged particular issues that required more in depth research than initially thought.

One of the characteristics of exploratory research is that the research design is more flexible, open to change and able to accommodate developments that cannot be planned or foreseen because the field is relatively unknown. In this sense the literature review helped to reformulate and redefine the research questions and objectives to focus more on actual issues. According to (Punch 2005:46) the characteristics for good research questions are:

- "*Clear*- easily understood and unambiguous
- *Specific*- concepts are at a specific enough level to connect to data indicators
- *Answerable*- what data are required to answer them and how can the data be obtained
- *Interconnected*- related to each other in a meaningful way
- *Substantively relevant*- interesting and worthwhile questions"

Through the literature review it was shown that the research questions were clear, interconnected and substantively relevant. However, issues about the specificity of the concepts arouse. It was clear that the concepts needed to discuss the issues were not developed enough by previous research and that this would need to be addressed within the study itself. In this manner the data required to answer all the research questions could be obtained.

The main conceptual issues can be grouped into three main categories:

Content typology for digital resources:

- There is no agreed vocabulary for electronic resources and even less so for non-formal electronic resources. This increases the difficulty of addressing the issues of how these are affecting scholarly communication and publishing as the lack of vocabulary limits analytical discourse. For example, the term 'data', is used to address a wide variety of different electronic resources that have differing characteristics. This leads to difficulties when trying to identify the differences between a wide spectrum of dissimilar types of digital academic research output in the form of data because of the lack of a more sophisticated vocabulary. This is a key issue for research into the impact of electronic resources.

Content and repository objectives:

- There is a lack of models for evaluating the effectiveness of IRs, in particular for the ones that aim to manage an array of different types of content. Although repository objectives have been defined in the literature there is apparently a disjointed discourse between the repository objectives and the types of materials that they collect. A need was discovered to examine in more detail the

repository objectives in conjunction with their content collection policies and actual content.

Content and depositors:

- Although there is a documented awareness of the lack of self-archiving, there appears to be little evidence of how material within the repositories is actually being deposited. The way that scholarly communication and publishing will be affected in the long run by IRs will be determined by who is depositing and what type of material they are depositing. However, there seems to be little awareness of who is currently depositing and no insight into the role of repository staff in this procedure. Repository staff depositing behaviour has been largely overlooked by the literature and the deposit work-flow processes in IRs are unclear. It was deemed necessary to address a specific question to this aspect.

Table 4 shows the initial set of research questions and the adjustments made following the literature review. It is important to point out that the definite research objectives and questions are presented in the Introduction chapter. This section aims to illustrate the reiterative process through which these were defined.

Objective	Initial question	Added questions
Examine repository approaches towards collecting and disseminating non-formal electronic resources	To what extent are electronic resources in repositories being used?	How are non-formal electronic resources managed in institutional repositories? What are the attitudes towards non-formal electronic resources amongst repository managers?
Investigate range and distribution of non-formal electronic resources and current typology limitations		What are the different types of non-formal electronic resources within institutional repositories? What is the distribution of non-formal electronic resources within repositories?

Explore new methodological approaches for measuring and evaluating use of electronic resources	What methodology can be used to evaluate the use of electronic resources in repositories?
Discuss implications for informal scholarly communication and publishing	What are the implications in the production and use electronic resources in repositories for scholarly communication and publishing? How can the boundaries between formal and informal publishing boundaries be defined in a new repository-based environment?

Table 4 - Reiterative design of research questions after literature review

The literature review was used to ascertain the key concepts that led to the identification of the variables related to the concepts to be studied. These variables acted as indicators and thus "when pursuing research objectives, we focus our attention on certain variables, observing them either to see how they appear or how they change" (Bouma 2000:49)

The variables for this research were identified as:

- **typology** of electronic resources
- **distribution** of electronic resources
- **attitudes** towards non-formal electronic resources
- **collection** policies
- **work flow** processes
- repository **objectives** and drivers
- **visibility** of resources
- **usage** of resources

These key concepts from the literature review were used to inform the design of the remaining part of the research methods.

Online questionnaire

Questionnaires[25] are an effective and popular research method of gathering information from a large population and have been used frequently in LIS studies, especially for surveying user needs and evaluating services (Williamson 2002:237). They are relatively easy and quick to administer and can provide a combination of quantitative and qualitative data if a combination of closed and open-ended questions are used. They are particularly useful for benchmarking and gaining an overview of a particular scenario and they provide timely and rapid results. Additionally online questionnaires allow for a worldwide sample at relatively low cost. In addition, the data are already in digital format that allows accurate and fast quantitative analysis using software programs.

One important drawback with online questionnaires for some population samples, is that an online survey may only be answered by respondents with an Internet connection (Burke and James 2006:19). Even for those with access to the Internet, online questionnaires tend to favour people who are familiar with the online world and are willing to use this method. However, as online surveys have become better known and more people are online, these factors have become less important. In the LIS field, online questionnaires are frequently used to research information seeking behaviour online with work on digital libraries, email discussion lists, use of the internet and others (Williamson 2002).

Online questionnaires are a particular popular research method and it could be argued that with the increased availability of easy to use software for designing them, there is a false perception that they are easy to administer (Burke and James 2006). However, it is important to design

[25] The research uses the terms questionnaire and survey interchangeably although as pointed out by Williamson strictly speaking they are different. A survey is a research design and the questionnaire its most common instrument.

online questionnaires carefully so that the data collected are relevant to the research objectives. According to (Tanner 2002:92-93) typical stages for planning and conducting a survey are:

- A general idea of the area to be investigated is derived from the literature search and then narrowed appropriately to ensure the project is manageable.
- Aims and objectives or research questions to be addressed by the questionnaire are devised. This ensures that the questionnaire questions are focused.
- Target group for survey is defined as well as whether the entire population can be surveyed or if a sample will be targeted.
- Survey technique (mail or email questionnaire, face-to-face interview, telephone interview) is determined.
- The survey instrument (questionnaire or interview guide) is written and pilot tested. Necessary revisions are made.
- A letter or email is written that explains the nature and aims of the survey to target population to enlist participation.
- Survey is conducted. If necessary follow up invitations calls for participation are done
- Data collected, recorded, analysed and interpreted.

The initial goal of this data collection stage was to contact a large number of repository managers and benchmark their opinions and knowledge about a range of issues with regard to the repositories they administrate. An online survey was considered the best research method for this goal. Simply due to time and geographic constraints a paper and pen questionnaire was not feasible or desirable. An online questionnaire would allow for a worldwide sample. In addition, due to the nature of their work, repository managers would be expected to be familiar with Internet technologies and online surveys and it was therefore reasonable to expect the target population to react favourably to this method.

Aims and objectives

The main aim of the questionnaire was to gain a better understanding of repository content typology and use in relation to repository depositing practices. The main objectives were:

- To identify key players involved in defining repository content … *who decides what goes into a repository?*
- To benchmark repository administrators' views on the value of different content types … *what is useful in the repository?*
- To benchmark administrators' views on the function(s) of a repository … *what is this repository for?*
- To build a picture of the administrator's perception of the use of the materials within the repository … *are the materials used? What for?*
- To gain a better insight into depositing patterns … *Who deposits what?*

Another additional objective was to recruit repository participants for the case studies, which will be discussed afterwards.

Designing the survey instrument

Previous surveys for repositories were reviewed in order to help in the design of the questionnaire with the two-fold purpose of examining the language and terminology used and to learn from previous pitfalls and difficulties encountered. As mentioned in the literature review the repository field currently contains numerous terminology inconsistencies and it was considered important to build on previous work in order to contribute towards stabilization and to make results comparable. A few surveys for repositories have been done (Lynch and Lippincott 2005; Westrienen van and Lynch 2005; Rieh, Markey et al. 2007) and these were examined. The findings of these surveys have already been detailed in the literature review and so in this section we will focus on their design of the survey instrument.

The 2005 census of academic institutional repositories (Westrienen van and Lynch 2005) was developed using data collected for the joint conference between the Coalition for Networked Information (CNI), the Joint Information Systems Committee (JISC) and the SURF Foundation

on "Making the Strategic Case for IRs". The organizers solicited data on the current state of IR deployment from 13 nations using a questionnaire. The main objective of this survey was to gather information about repositories on a national level. The respondents were asked to provide data about the number of repositories in the country, average number of documents, coverage in terms of types of materials, academic participation, as well as several aspects of national policy. Although the survey was useful to identify common factors of interest such as the coverage in terms of types of materials, the actual questions were directed at a national rather than an individual repository level. The survey was quite quantitatively focused requiring respondents to give their replies as estimated amounts or percentages.

The study reports several difficulties with their data collection procedure. One of the main problems was detecting possible national data sources for information on IRs. It was not clear what organization(s) should be approached that could provide this information and the study hints at disagreements between the figures presented. Several nations could not provide all the data requested in the questionnaire. There were also problems with scoping and interpretation. In some cases where the IR framework is more unified, such as the Netherlands, the UK and the US, they reported that additional interpretation was provided and this helped to clarify the data provided. Additionally there was confusion with the terms employed in the questionnaire. For example, *records* is interpreted differently in different nations. In the USA it is generally assumed that a record will include the digital object while in the Netherlands records can be metadata only. This survey was probably the first effort to gather comparative international data and in this sense is an important start. However, the way in which the data were actually collected and how each nation calculated the percentages is not detailed. The sample population for each nation is unknown and it is therefore unclear if the data supplied were from an individual or from a larger sample of IRs.

The exception is the data for the US which, although did not attempt to survey a statistical example, did at least report in a separate article (Lynch and Lippincott 2005) on how the data

was collected. The questionnaire was emailed to 124 HE institutions in the US that belong to the CNI, one of the co-organizers of the above mentioned conference. They also point out that they did not attempt to define IR but rather asked the respondents to give their own definition. One of their findings was the "confusing relationships at many institutions among digital libraries, digital research collections and collections of materials in IRs" (Lynch and Lippincott 2005:unpaginated) based on the types of materials that they were collecting. They point this out as an area that requires careful future analysis.

The other survey examined was a census of institutional repositories in the US comparing institutions at different stages of IR development (Rieh, Markey et al. 2007) discussing five key components: leaders, funding, content, contributors and systems. This survey was particularly focused on gathering data from all universities regarding IR. They compared the responses for four categories: No planning to date, Planning only to date, Planning and pilot-testing and Public implementation of IR. The survey was directed at library directors regardless of whether they had a repository or not. Interestingly they noted that with respondents that had no repository or were currently only planning a repository, it was more likely that the library director would respond to the survey, whilst in the case of respondents with an implemented repository it was more likely to be answered by a repository manger or similar role. This helped show the viability of looking directly for repository managers to answer the online questionnaire for this survey as we were looking for information from implemented IRs.

From looking at the previous IR surveys it was decided that a combination of closed and open-ended questions would reliably pick up on the issues. As mentioned previously, the repository landscape is currently a contentious one and there are a considerable amount of expectations on the future of scholarly communication and publishing that depend, to a large degree, on the success or failure of repositories. It was therefore considered a good idea to base the survey on a significant number of closed questions. However, repository administrators' opinions and views are part of the objectives of this survey so open-ended questions were also included. This also

allowed repository managers to point out additional issues that were not addressed in the questionnaire but are also important.

Questionnaire content

The survey consisted of eight sections: introduction, repository information, repository materials, repository deposits (a and b), repository function and use, repository administrator information and closing section.

The introduction presented the objectives of the survey, details on issues about confidentiality and a link to further information about the project. It also included information about the estimated time that the questionnaire would take to complete and an email address for any questions. The text was designed to be concise but informative in order to interest respondents and encourage their participation.

The second section - repository information- gathered details about the manager's repository including name, URL, name of university or organization hosting the repository and country. Respondents were also asked about the stage of development, age and number of items. This idea follows on from the 2007 US survey of IR deployment (Rieh, Markey et al. 2007) that organizes responses according to different categories ranging from no IR and no plans to have one up to IR implemented. In this particular case the data we were collecting were on implemented IRs but we asked managers to classify their repository in one of the following categories: prototype, recently launched/initial stage, fully operational repository or other. In a similar vein to the 13 nations census (Lynch and Lippincott 2005; Westrienen van and Lynch 2005) repository managers were asked how many items (equivalent of records) were currently stored in the repository. Although the difficulties of this measurement were known, there is currently no alternative or standard form of counting[26] and although flawed it does give some

[26] The US study suggests measuring IRs in terms of mass storage as a possible alternative option. As with item number this method is also problematic but could give an indicator of more non-formal electronic resources being present. Metadata records and eprints

context. Additionally size was not one of our key variables, and so together with the age of the repository, this information was addressed more as demographic contextual information than as data for analysis.

Section III -repository materials- gathered information about the types of electronic resources that the repository stores. Managers were requested information about the types of materials that they collected within the repository. In order to make data comparable and attempt to categorize material types respondents were expected to select from a list. Defining this list of materials was an initial challenge and again the previous IR surveys were consulted to see how these categories had been defined and the list compiled.

The 13 nations survey (Westrienen van and Lynch 2005) asked nations to indicate the percentage of different content types from a list of six different types. These were: Articles, Theses, Books, Primary data, Video Music etc., Course material and Other. The study does not offer any explanation on how the list was compiled. The study with the US data on the other hand (Lynch and Lippincott 2005) offers a much longer list with over 30 different types. According to the study it was designed using the initial 13 nations survey list and other types were added based on their own insight. The 2007 US survey (Rieh, Markey et al. 2007) indicates that their questionnaire listed "three dozen digital document types" but no further details are given. For this study it was considered that the 13 nations survey list was too short with only six options available but the list for the US data, with over thirty types was too exhaustive.

as pdfs take up relatively little space compared to other formats such as images, videos and so forth. So for example, a low item count but high storage could indicate the presence of non-formal materials whilst a high item count but low storage would probably refer to a metadata record only IR. For this study however, this information was not requested as this is information that is usually only handled by US repositories.

An alternative was found using OpenDOAR[27], a directory of Open Access Repositories that aims to be authoritative with good quality information about each repository as this is compiled by OpenDOAR staff rather than automatically. Information includes the type of content that a repository keeps and their content type listed consisted of 15 options. It was decided that this content typology would be used. The final options were:

- Administrative documents
- Books and book chapters
- Conference proceedings
- Datasets and databases
- Images, maps, diagrams
- Learning objects
- Audio-visual materials and multimedia
- Patents
- Postprints
- Preprints
- References/bibliographies
- Software
- Theses and dissertations
- Reports
- Working papers

This content list was designed to serve as a starting point as in the next question respondents were asked to list all the types of electronic resources allowed for deposit that were not mentioned above. We then gathered data on the frequencies of the content types.

The following section - repository deposits- occupied two sections, IV and V, and collected information relating to what groups made decisions about the types of materials deposited in the repository and what groups were depositing what types of materials. As closed questions were being used in order to facilitate data comparison, we needed to define the different possible

[27] OpenDOAR is reviewed extensively together with other repository listing, ROAR and OAIster, in the *Overview of repository registers and growth of repositories* of this chapter.

decision makers and depositing groups. These are presented in Table 5. Data on deposit frequency by group were also recorded.

Decision makers groups on deposits	Depositor groups
Myself (repository manager)	Lecturers/Researchers
Library	Library staff
Special committee	Administrative assistants
Computing services	Students
Other	Other
Unknown	I don't know

Table 5 - Decision makers for deposits and depositor groups

Section VI -repository function and use- gathered information about the importance of different objectives of a repository. Respondents were asked to indicate how relevant a particular repository function was in relation to their repository objectives as well as their level of agreement with several key concepts that had been identified in the literature review as contentious such as 'repositories should only contain peer-reviewed resources´.

The section also collected data about the use of the resources within the repository and the methods for measurement of use. Repository managers were to indicate on a predetermined scale how often the resources within their repository are used, who is in charge of measuring this and how this is done.

This section finished with an open-ended question inviting respondents to offer their general views or concerns about the types of electronic resources deposited in repositories and the different groups who deposit. This was important as it allowed repository managers to comment freely on particular aspects from the closed questions or indicate additional aspects of IRs that had not been addressed in the questionnaire. The idea of this section was to allow input and dialogue from the respondents.

Section VII requested information about the respondents (name, job position and email). This information is not available in the results and was collected for administrative purposes only. Respondents were also asked if they would be willing to be contacted by email or by phone for further discussion about their replies. The final section thanks the participants and also offered them the option of receiving the results of the survey.

See Annex 1 for a complete copy of the survey.

Piloting the questionnaire

The survey was piloted with four repository administrators. One particular section, V regarding depositing behaviour according to user groups, posed some difficulties and was amended accordingly. This section asks respondents about the differences in depositing behaviour when done by the author of the electronic resource and when deposited by a third party. This proved to be a somewhat unusual concept and caused a certain amount of confusion. It may be that repository managers are currently more concerned about gathering content by whatever means and have not thought too much about the differences in content type between author and third party depositors. However, this is obviously an important issue as shown by the great concern in author's lack of motivation for self-depositing. Revisions and amendments to the questionnaire were done from the pilot comments. The questionnaire was also submitted to supervisors for final remarks and approval.

Questionnaire language

Non-English speaking countries or countries where English is not a common second language can be under represented in a global survey of this kind. Because of this the survey was translated and offered also in Spanish. It would have been useful to have more languages available (for example Portuguese, Japanese or Chinese) but unfortunately time and knowledge constraints did not allow this. Links to both versions of the survey were available from the

email invitation where applicable. Survey replies in Spanish were translated and merged with the English results.

Online questionnaire design

The questionnaire was designed and put online using Survey Monkey, an online survey facility. It allows an unlimited amount of responses per survey as well as producing summary and detailed reports on findings and the facility to download responses in various formats for further analysis. Figure 4 shows a screen shot of the survey.

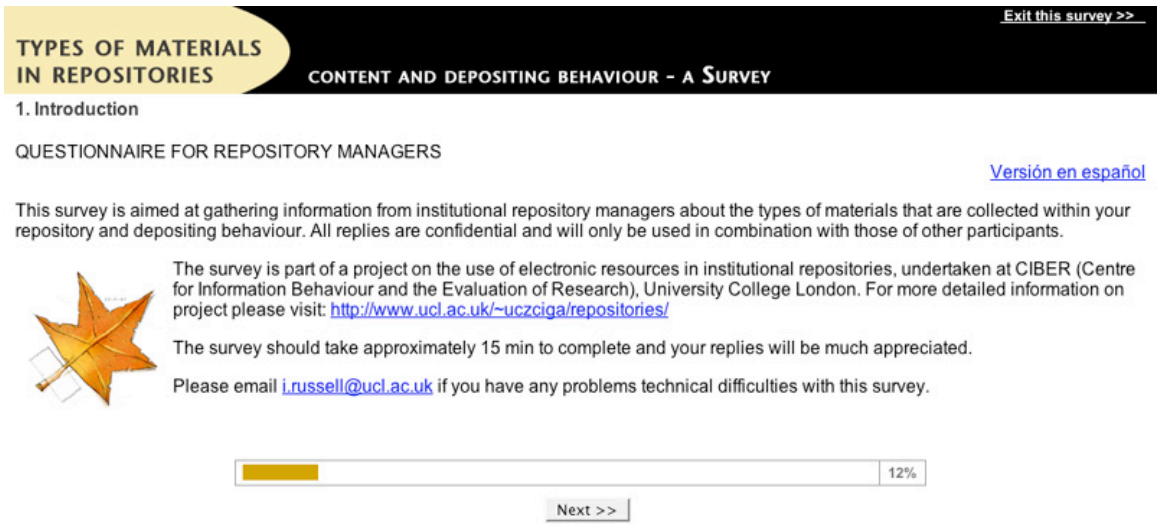


Figure 4- Screenshot of survey

Non-cooperation is one of the biggest issues for online questionnaires (Tanner 2002) and frequently cited reasons for this are surveys that appear to be poorly conceived or are unnecessarily long or complex. Extreme care was taken to provide adequate but concise information about the survey throughout the different sections. The layout and design of the survey, including text size and colour, were designed taking into consideration basic usability web issues.

The survey was opened on the 23rd of July 2007 and ran until the 18th of September 2007.

Sampling

Making sure that the population that replies to the survey is representative of the whole target population is one of the key issues for questionnaires. One of the great drawbacks of online questionnaires is a poor response rate (Tanner 2002) especially with the increased use of this method by researchers and also postgraduate students. A high response rate is important if results are to be generalized accurately from the sample to the general population.

Identifying the total population of repository managers was the first stage to determine response rate. There are currently no lists of worldwide repository managers. However, it seems fair to assume that a repository manager would manage at least one repository. OpenDOAR[28] was used to determine the number of repositories worldwide. At the time of the study (July 2007) there were 927 repositories registered with OpenDOAR from a range of countries, as shown in Figure 5.

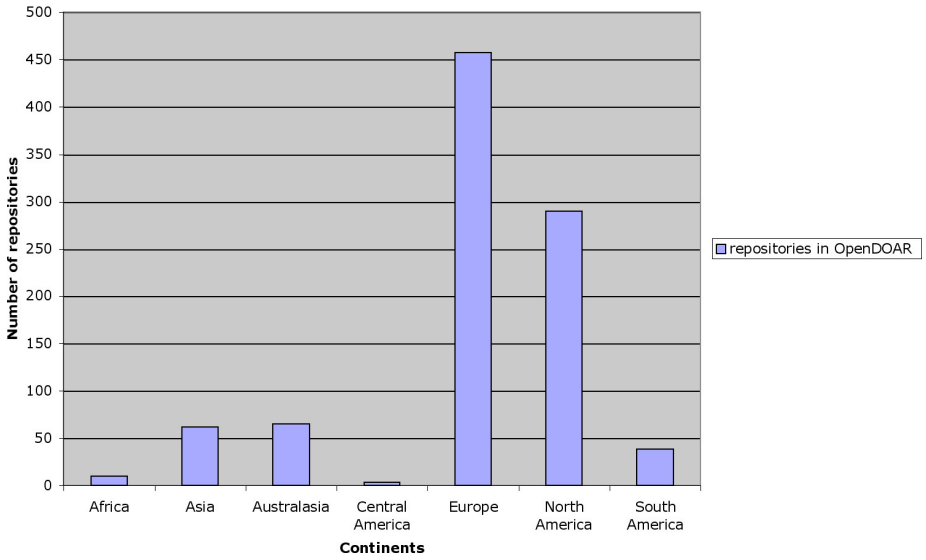


Figure 5 - Repositories in OpenDOAR by continent

Repository managers make up a relatively new group of professionals and it is only recently that posts have been set up specifically with that job title and description. However, repository

[28] OpenDOAR is reviewed in the *Overview of repository registers and growth of repositories* of this chapter.

managers tend to come from more established groups of professionals that work in the information science and library world, such as digital librarians, information professionals, computer engineers and other related groups. These groups have already established a number of important communication channels, such as the ones described below. These were used in order to send out an invitation to repository managers to answer the online questionnaire.

The particular target group for the survey, repository managers was indicated in the subject line 'Looking for repository administrators' of the invitation email that was sent out. This email was sent out using seven pertinent discussion lists. The discussion lists were selected according to their subject matter and where there was a reasonable expectation that repository administrators would be subscribed. These were OpenDOAR, Dspace-General, UKCORR (UK Councils of Research Repositories), JISC-Repositories, JISC- CETIS Metadata and Digital Repositories, CODATA (Committee on Data for Science and Technology of the International Council for Science) and SIGMetrics. These email lists cover the following main topics: repositories general, advocacy, social aspects of repositories, metadata, infometrics, data repositories, non-technical aspects of running a repository. The first five are dedicated discussion list for repositories with a high likelihood of overlap between subscribers. See Annex 2 for a more in-depth description. The initial invitation emails were sent out between the 23rd and the 30th of July 2007.

A preliminary analysis was done on data downloaded on the 16th of August 2007. The report focused on the data for the first four questions: Country of hosting university or organization of repository, Classification of repository in terms of development, age of repository and number of items contained. From these initial results further targeted email invitations were sent out focusing on repositories from countries that were underrepresented in the survey and which could not be explained by language barriers. The first invitation focused on repositories in India, whilst the second focused on the Latin American region. Invitation texts can be found in Annex 3.

The 7 email lists had a combined total of 4266 subscribers, plus 55 targeted email for a total of 4321 email invitations. Table 6 shows a breakdown of number of subscribers by distribution list, bounce rate and likelihood of repository managers being subscribed.

Mailing list name	Num of subscribers	Coverage	Likelihood of repository administrators subscribers	Bounced
OpenDOAR	1117	International	High	100
Dspace	1115	International	High	Unknown
UKCORR	80	UK	High	Unknown
JISC Repositories	920	UK	High	Unknown
CETIS Metadata and Digital Repository	303	UK	High	Unknown
CODATA	214	International	Low	Unknown
Sigmetrics	517	International	Low	Unknown
India (targeted)	19	India	High	2
Latin America (targeted)	36	Latinamerican region (Brazil, Chile, Colombia, Mexico, Peru, Venezuela)	High	2
TOTAL NUMBER	4321			

Table 6 - Email list, number of subscribers, likelihood of repository administrator subscribers and bounce rate

Completion and response rate

Completion rate

Data were downloaded from Survey Monkey on the 18th of September 2007.

The Survey Monkey system reports 186 responses of which 150 completed the survey. This is just over an 80% completion rate. Survey Monkey defines completed responses as those who press the 'Done' button at the end of the survey.

The 186 responses were downloaded and analyzed using SPSS 14.0. The findings were slightly different. Only 16 responses were eliminated as they were empty and contained no data (only the initial IP address which means that the survey was viewed but no responses were given). The remaining 170 responses, and although a few contained some empty responses (no questions were obligatory), contained enough data to be useful.

Response rate

Calculating response rate was complex. The survey was sent out to a total of 6 email lists, with a combined total of 6495 subscribers, plus 58 targeted emails for a total of 6,553. This gives a response rate of 2.59%. However, this figure is misleading for a number of reasons:

- Bounce rates: bounce rates were only available for one email list (OpenDOAR). Of 1117 email addresses, approximately 100 emails (8.9%) were bounced back. This figure is not available for the other emails lists and could therefore not be calculated. Of the targeted invitations, 58, four emails were returned (6.8%).
- Duplicate email addresses: considering the subject matter of the email list discussions it is quite likely that the same email addresses may be found in the different lists. For example, the targeted invitation emails were taken from the OpenDOAR registry, which was already covered by the OpenDOAR email service.
- Repository administrators- The survey specifically requested repository administrators (ie. The subject line was *Looking for repository administrators*) and although some of the subscribers have this or similar position, they are a fraction of the total. It is not possible to calculate this value.

It was therefore decided to estimate response rate based on the total number of repositories currently registered in the world in order to assess the value of the sample covered in the survey. It is fair to assume that each repository will have one repository administrator (or similar role)

and indeed it is also possible that one person may administer more than one repository (in particular the cases where one institution has numerous repositories).

The total number of repositories per directory listing of the three main repository listings (OpenDOAR, ROAR and OAIster) was determined[29] and the response rate calculated. The average number of repositories according to the directories is 875 worldwide, which gives an estimated response rate of 19.42%.

Directory listing	Total number of repositories	Response rate
OpenDOAR	927	18.33%
ROAR	844	20.14%
OAIster	854	19.90%
Average	875	19.42%

Table 7 - Response rate according to total number of repositories

Case studies

The online survey data was also used to select potential candidates for the next step in the research, seven case studies of repositories. Case study however, is a difficult concept to define. Gerring mentions eight different definitions in "What is a case study? The Problem of Definition" (Gerring 2007:17). In his working definition he also makes a distinction between case study and cross-case: "A *case study* may be understood as the intensive study of a single case where the purpose of that study is – at least in part – to shed light on a larger class of cases (a population). *Case study research* may incorporate several cases, that is, multiple case studies. However, at a certain point it will no longer be possible to investigate those cases intensively. At the point where the emphasis of a study shifts from the individual case to a

[29] Numbers taken on the 14th of August 2007.

sample of cases, we shall say a study is a *cross-case*" (Gerring 2007:20). This is similar to the concept of collective case study where a number of cases are studied in order to investigate some general phenomenon (Silverman 2005). Following these definitions the case studies for this research are indeed cross-case or collective case study.

Gerring also points out that an implication of the term case-study is that the unit or units in this case, are not necessarily representative of the population as unit homogeneity across the sample and the population can not be assured. So although the case studies are institutional repositories, it is not assumed that the results from our study will necessarily produce determinant characteristics of IRs. It may well be that they may offer specific examples of the different approaches and definitions of repository managers and institutional repositories. These case studies were designed to clarify issues that were raised from the literature review and the online questionnaire regarding repository content typology, depositing activity and motivation, content use and visibility.

Selection of case studies

In the online survey respondents were asked if they would be willing to provide further information about their repository for the study. The repositories that declined were removed, leaving 90 repository managers that would be willing to be contacted. These 90 repositories were looked at specifically, together with the answers that they had provided in the survey and detailed notes were taken on repository content type, number of items, age and country, as well as the openness and detail of the repository manager's open-ended responses. Figure 6 shows a breakdown of these repositories by country.

Country	Agreed to interview
Australia	3
Brazil	5
Canada	3
Croatia	1
France	2
Germany	2
India	4
Italy	5
Mexico	11
Mongolia	1
Netherlands	1
Norway	1
Poland	3
Portugal	1
South Africa	1
Spain	2
Sweden	1
Switzerland	2
UK	14
Ukraine	1
USA	26
TOTAL	90

Figure 6- Repositories for selection by country

Initially it was hoped that apparent categories of repositories would appear from the analysis but the repository landscape is currently very heterogeneous and no obvious groups appeared. It was considered that another option would be to select the case studies according to other criteria, such as by degree of development, age or country. Neither degree of development nor the age seemed particularly useful as using well established and older repositories could skew the results and possibly portray a more advanced scenario than what actually exists. In general very new or starting up repositories tended to have little content, which could also impact negatively on the study. It was therefore decided that the best option was to select a country and sample their particular case studies. Countries with the largest representations, the USA and the UK, were chosen which allowed for larger scope in the selection of appropriate repositories (see Figure 6). In addition, both countries are well advanced in the repository environment and are considered leaders in this field[30]. The present study is being done in the UK and therefore in

[30] Both the Netherlands and Germany are also fairly advanced as well. However, language barriers were considered an obstacle.

order to facilitate interviewing, the best option was to select repositories from this particular region. This also placed less pressure on financial resources and time constraints.

There were a total of 13 responses from the UK from a total of 12 universities (one university had two repositories). Initially, some repositories were eliminated either because they were too small and did not have sufficient content, were still in a prototype stage or the content type was not diversified enough (ie they were mainly eprints). In total there were seven repositories remaining from six different universities. This was considered a suitable sample size for case studies given the time and financial constraints. In addition, this part of the research was designed to be more in depth rather than a large cross case study. The final selection shows variety in terms of size and stage of development and therefore covers a range of perspectives. Table 8 shows the repositories by number of items and stage of development.

AGE	REP	ITEMS	REP	TYPE	REP
< than 1 year	1	101-500	3	Recently launched/initial stage	2
1-2 years	1	1001-5000	2	Fully operational repository	5
2-3 years	2	10,001-20,000	1		
3-4 years	1	over 100,000	1		
4-5 years	1				
> 5 years	1				
TOTAL	7		7		7

Table 8- Selection of repositories by age, number of items and type

Case study data collection

Data collection methods used for the case studies were in-depth interviews and link analysis. These are outlined in the following section.

Interviews

The online survey provided data about repository manager's attitudes towards different types of materials and their repository content policies and objectives. In order to find gather more in-

depth information about issues such as the type of relationship between repository objectives and policies or the work flow processes in place, interviews were conducted with the repository managers from the case studies.

Interviews are a research method that is particularly adequate for exploring processes both social and political (Rubin and Rubin 2005) and are appropriate to interpretivist methods (Williamson 2002:242). Interviews allow for complex and complete responses and explanation, and clarification can be provided to the respondents as well as to the interviewer. In this way interviews can aid with concept clarification by allowing the researcher to directly engage with the interviewees' definitions (Williamson 2002). One of the problems with self-administered questionnaires is that definitions are either pre-imposed to prevent confusion or are left open leading to uncertainty about how the concept was understood by the respondent. For example, the survey of IRs in 13 nations (Westrienen van and Lynch 2005), reports that a lot of the data they collected from the survey was clarified with additional interpretations by talking to the respondents. So while the data collected from the online questionnaire were used to build up a an overview of the repositories on a global level, the interviews with repository managers sought to detail and find answers to important points that were detected from the questionnaire. The interview scripts were written after the questionnaire data had been analyzed in order to examine in further detail subjects that had come up from their replies.

It was decided that the best approach to the interviews was a semi-structured, open-ended and relatively informal design in order to encourage repository managers to discuss specific themes and topics issues at length, including personal experiences, thoughts and ideas (Rapley 2004). It also allowed the interviewer to follow up any particular leads or issues that arose. However, it was important to use an interview script in order to ensure similar coverage of data collection for variables across all case studies.

Interview script

The repository administrator interviews main aims were to gain a deeper understanding of repository development, content ingestion work flows, depositing behaviour, content typology, resource usage monitoring, dissemination and future.

The objectives were to:

- To build a historical framework of repository development… *how did the repository develop?*
- To describe content intake workflows for the repository… *what mechanisms are in place to put electronic resources into the repository?*
- To gain a better understanding of depositing practices… *who deposits what and how?*
- To better understand the makeup of repositories in term of content… *what types of resources are in repositories and in what quantities?*
- To identify types of resource usage monitoring… *How are repositories monitoring the use of their resources?*
- To benchmark repository administrator's views on the use of electronic resources in the repository… *Are electronic resources in repositories used (and what for)?*
- To identify repository administrator's strategies for disseminating the repository and its contents?… *What do they do to disseminate the repository?*

The interview script was divided into five main sections: background, content intake workflows, depositing behaviour, usage and the future of the IR. After the script was written it was sent to both supervisors for comments. The interview was piloted with two repository administrators and amended accordingly.

The final interview guide is presented as Annex 5.

Interview procedures

The repository manager from each one of the case studies was contacted and asked for an interview. Interviews were carried out over a one-month period from February 2nd to March 3rd 2008. Approximately two interviews were carried out per week. In all cases interviews took place at the repository manager's workplace. In two cases interviews were carried out with more than one interviewee.

Interviewing professionals about their work with repositories did not present any major ethical or neutrality issues, although normal ethical procedures were taken, such as asking for the interviewees to sign a consent form and indicating that they would be recorded. Additionally, interviewees were informed that all case study data would be anonymized. Interviewees were also informed that in a situation where a comment was only comprehensible if associated with the name of a given repository, the interviewee would be contacted and asked permission for his/her name to be used and would be given the opportunity to check the accuracy of the quotation. Interviewees could also request that certain information remain confidential and not for publication. A copy of the consent form is available as Annex 5.

Although the original intention was for interviews to last no more than thirty minutes in all cases they took between fifty and sixty minutes. Interviewees were very willing to talk extensively about the repository and its development. The interviews were recorded and a few notes were taken. On two occasions a repository manager requested that the recorder be turned off, in one case to discuss particular delicate political affairs and in the second, because she did not feel qualified to go on record talking about a particular aspect of repository development. This does not mean however, that recording the interview somehow made the data gathered less 'real'. The interviewees were asked to respond to their questions within their role as repository managers and asking the tape to be turned off was probably due to wanting to answer or

comment something outside this role. "Such off-tape talk is not somehow more 'authentic', it does different work, it emerges from and reflexively creates a different context (…) Importantly it documents that the prior talk was the product of a specific interactional context (and a specific identity) and that now the context (and the identity) has shifted." (Rapley 2004:19).

Interview analysis

Interviews were transcribed in order to facilitate analysis. Transcribing them myself was seen as an important way to begin the analytical work, notes were taken and certain themes began to emerge. "In this way, I got to repeatedly listen to the tapes, and so generate, check and refine my analytical hunches whilst simultaneously producing a *textual* version of the interaction that could be used for further analysis and reports" (Rapley 2004:27). The transcription process provided an initial general analysis.

These transcriptions were then printed and analysed further. Due to the particular combination of qualitative and quantitative research methods used for the study, it was considered that grounded theory approach would be useful for analysing the interview data. Grounded theory has been successfully employed in LIS research in particular with information seeking behaviour studies (Mansourian 2006:396), although it has also been used for online learning, user studies and classification. Grounded theory is a research approach that has its origins in social research and therefore fits in well with the socio informatic focus of this thesis.

Grounded theory was originated by Glaser and Strauss in the 1960's although since then it has evolved and there are currently several interpretations (Dey 2004). The differences for this study are not particularly relevant as a general definition of grounded theory suffices. "Grounded theory is a general methodology for developing theory that is grounded in data systematically gathered and analyzed. Theory evolves during actual research, and it does this through continuous interplay between analysis and data collection" (Strauss and Corbin 1994

cited in Mansourian 2006:387). The grounded theory approach works by codifying the data into categories for comparison. An open coding system was used in such a way that the codes were determined from the data analysis rather than by a preconceived set imposed on the interviews (Kim 2000).

Given the relative novelty of interviewing repository managers and depositors, it was considered pertinent to uncover the important issues through the analysis rather than trying to determine if preconceived issues were perceptible. Although the script was designed to steer the interviews, many important themes and strands were found to weave across the entire conversation with individual repository managers, as well as appearing in the dialogue with other repository managers. The coding system would effectively allow the comparison of themes across the different case studies (Charmaz 2006). The results are presented in the following chapter.

Measuring the use of electronic resources

One of the main objectives of the thesis and defined as key issues, are the visibility and usage of electronic resources within IRs. These concerns could be only partially addressed with data from the survey and the interviews so an additional specific research method would be needed to enhance the existing data and contribute to further addressing the main research objectives. Selecting and applying a research method to collect and analyse this data is both a way of attempting to answer the question of use and visibility, as well as a way of evaluating the effectiveness of methods for measuring electronic resource usage. Two methods were reviewed and applied wherever feasible: log analysis and link analysis. Both methods are described below.

Log analysis

The use of transaction server log analysis is becomingly increasingly popular as a way of measuring and evaluating the use of electronic resources. Although there are some

disadvantages, the analysis of logs can be considered an important first methodological step (Nicholas, Huntington et al. 2005b). The advantage of logs is that they offer large and fairly robust data sets on the use of electronic resources. As pointed out by Nicholas *et. al.*, they record everyone that uses the system, so there is no need for sampling, and more importantly, they are an immediate and direct record of what people have done, "not what they say they might, or would, do; not what they were prompted to say; nor what they thought they did" (Nicholas, Huntington et al. 2005b:251). Although it is difficult to reach any definite conclusions about usage from log analysis alone, when combined with other information (such as demographic information about the users) or qualitative interviews, deep log analysis may provide important insight into, as in this case, the use of IRs.

Serious attempts were made to gather log information for analysis from all seven case studies and to glean initial information about institutional repository usage. However, this proved slightly problematic, as some repositories were either unwilling or unable to provide these data. The reasons were severalfold: data protection concerns with providing usage information with IP addresses, (unless of course it was anonymised which would require extra work for the server administrator and would also limit the log analysis); changes in repository servers which make log data coverage erratic or again it would require asking server administrators to do extra work to provide the logs and some repository managers were reluctant to do this. As repository managers had already been very generous with their time in agreeing to the interviews and responding to the surveys, it was considered inappropriate to insist. Some of the repositories that did have the logs available reported different formats ranging from raw server files to usage information condensed and provided by a software package. This meant that comparing log information from different servers would be complicated and time consuming. This experience confirms reports from the literature on the difficulties associated with acquiring log data from repositories (Warwick, Terras et al. 2008). Finally, unless log data could be obtained from all the repositories, the log analysis for the case studies would be partial making comparisons between repositories more difficult.

For these reasons, formal log analysis was not carried out for the case studies, although this is an interesting future area for research. However, both in the survey and during the interviews, repository managers were asked about if and how they monitored the usage of the resources within their repository and about the usefulness of this information. In particular, during the interviews a relatively long period was dedicated to this subject. So although no systematic log analysis was done, relevant data related to monitoring usage through logs were collected.

Link analysis

Link analysis is a methodological approach for looking at web-related phenomena that is based on the premise that studying links to a particular resource can reveal important information about its perceived importance or usage, specifically it is related to a range of informal types of scholarly communication (Wilkinson, Harries et al. 2003). One of the advantages of link analysis is that the data can be collected directly from the web using commercial search engines. So although we could not look at usage statistics of electronic resources within the repositories, it was possible to look at what web pages were linking to resources within the repository. This could shed light on what electronic resources within IRs are being used for by analyzing what types of web pages were linking to them.

This section describes the methods used for the link analysis study applied to all seven case study repositories. Due to the nature of the research it was considered appropriate for the link analysis to have a more qualitative rather than quantitative focus. The main aim of the link analysis was to discover if and what resources within IRs are linked to (target page) and from what type of web pages (sources page). The purpose of this study was not to determine the amount of usage through link analysis (i.e. by counting number of links to a repository and implying a level of use) but rather by attempting to understand distinct usage of different types of items within a repository by looking at linkage. If novel forms of scholarly communication and publishing are to be achieved through new digital genres then a study of links to the

different types of resources available could provide important insight into if and how these changes are occurring.

The main objectives of the link analysis were to:

- To discover what types of resources are being linked to within repositories… *What are the links to?*
- To build a typology of source pages… *What different types of sources pages can we find?*
- To infer from source page typology, the use and perceived usefulness of resources within the repository… *Why are they linking?*

Collecting the links

The first step was to collect all the URLs of web pages (source pages) that link to a repository page (target page) from all seven case study repositories. Building a crawler for searching the whole web is not feasible, nor desirable and it has been accepted that link data collected from search engines are relatively reliable (Thelwall 2008). One of the main drawbacks is that no search engine covers the entire web so it is not possible to assume that all the links to a particular resource are included. This is an area of particular concern when using link analysis for more quantitative type studies. For example, if we wanted to compare the number of links to different items within an IR or even between different IRs. This however, is not a point of particular worry for this study, as it does not intend to be exhaustive by covering all links to a repository or determine absolute figures on usage. Rather the aim is to gather a selection of links to the case study repositories to develop an initial diagnosis of the different types of usage of repository materials found by looking at a sample of the links to repositories and to what types of resources. Link analysis studies can work with large data samples for a more quantitative approach (Thelwall and Harries 2004) or for a more interpretative analysis smaller link samples are appropriate.

Data link collection was done using Yahoo! search engine which currently offers this feature. Search engines vary in the search functions they offer. Google, Microsoft, Yahoo! and Altavista currently allow searches for all links to a single web page but only Yahoo! and Altavista permit searches to find all links to a particular domain, as required for this study. To overcome search engine limitations the use of poly representation has been suggested (Almind and Ingwersen 1997) by collecting search results from several search engines. However, since the data link collection did not intend to be exhaustive or comprehensive, this was not deemed necessary. The software used, LexiURL described below, allows searches on Yahoo! search engine which was appropriate for this study.

A search was performed to find links to the case studies but the repository homepage was deliberately not included. One of the objectives of this study is to determine specifically how much the resources within the repository are used rather than the repository as whole. The rationale was that this search strategy would collect links to the actual resources themselves rather than to the repository in general. The software used for all repositories builds similar URL structures that easily allow this kind of search. In addition, links within the repository or self-links (Björneborn and Ingwersen 2004; Thelwall, Vaughan et al. 2005), such as help pages and menu links, were also excluded from the search.

The exact search text was as follows:

linkdomain:repositoryurl.ac.uk -site:repository.ac.uk -link:http://repository.ac.uk/

Data were collected for each one of the seven case repositories on the 19th March 2008. LexiURL software, developed by the Statistical Cybermetrics Research Group[31] which has been used successfully for several link analyses, was employed. The software automatically runs multiple searches and then offers several tools for managing and analysing the data as lists of Web page URLs and hyperlinks. The software is freely available for download.

[31] Statistical Cybermetrics Research Group based at the University of Wolverhampton, see http://cybermetrics.wlv.ac.uk/index.html

The numbers of links per repository can be seen in Table 9. There was notable variation between the numbers of links per repository. As with any exploratory and novel method, it was difficult to determine what could be considered an adequate and reliable sample of links. Previous studies have either been large and quantitative (Thelwall and Harries 2004) or have used the article, rather than the link, as the unit of study (Kim 2000). The repository with the least amount of links had 22 in total whilst the largest had 989. The minimum, 22 was not considered adequate as a sample as it was too small compared to the number of links of other repositories. It was finally determined that at least one hundred links would be analysed initially for each repository, except in the two cases where there were less than that and all links were analysed. Once again, it was considered important to focus more on the nature of target and source pages of the links than on the actual number of links analysed.

Case study number	Number of links found
1	419
2	989
3	157
4	30
5	992
6	325
7	22
	2934

Table 9 - Case studies and the number of links

Two repositories had less than 100 links and so all links were analysed. In the case of repositories with more than one hundred, a random sample of one hundred links was analysed. Lexi URL alphabetically lists all links found automatically but also offers the option of links randomized by domain. This tool was used to obtain one hundred random links per repository.

Defining target and source page typology

Once the sample of links had been collected the next step was to look at the page from where the link was created (source page) and then follow the link and look at the page in the repository to where the link was directed (target page). Taking into consideration the qualitative focus on link analysis of the present study and the fact that the total sample of links was relatively small (a total of 552 links) each target and source page was visited and analysed individually in order to build a comprehensive target and source page typology.

Classifying web pages is currently still a contentious and complicated matter (Crowston and Williams 2000; Thelwall 2003b). One issue is the definition of the unit of analysis (i.e. what constitutes a web page?) and in particular when counting links, the degree to which it is necessary to aggregate documents. Most studies tend to take the basic HTML document as the unit of measure, but this is rather artificial as for example, a researcher's homepage can be one HTML page or can be spread over several HTML pages (a publications page, research interests, a CV) with links between them for navigation purposes. There are several proposals to resolve this issue (Thelwall 2002; Thelwall, Vaughan et al. 2005). For this particular study, as links were not counted, defining the units of analysis was simpler: target page and source page.

Genre identification of web pages is also still a problematic issue, especially because there are no standards and most genres are still under development (Crowston and Williams 2000; Rehm 2002) and physical characteristics which aid genre identification in print media, are not present in the digital environment, for example the page format of a newspaper, journal or book (Crowston and Williams 2000). Additionally, web pages or sites can effectively combine a number of genres further complicating classification (Cronin, Snyder et al. 1998; Crowston and Williams 2000). Although some work has been done towards automatic classification of web pages (Almind and Ingwersen 1997; Rehm 2002), it is still in development and unreliable. Most web classification studies, discussed below, have therefore taken a qualitative approach and examined each web page and developed genre typologies. Defining genre is particularly

meaningful for this study as it has important implications for certain aspects of scholarly communication and publishing. What are valid scholarly forms (genres) of publications on the web?

Target page typology

All repositories used a classification system for the items they store making it unnecessary to develop a new typology system for target pages. The initial step was to define the target type using the item type registered by the repository, such as conference proceeding, postprint, video or working paper. However, not all repositories used the same classification system, although there was a great deal of overlap and variations could be quite subtle. In some cases, the labels were slightly different but were conceptually the same as for example, article or journal article. In other cases, some repositories had a more granular definition whilst others used a more general label, as in for example, using the term Conference or Workshop item as opposed to Conference. A second step was to homogenize the different categories in order to be able to compare across repositories. An initial attempt to map these categories to the item types used for the survey was made but this proved unsuccessful due to the fact that the survey classification system was too specific for some of the repository classification systems. It was therefore necessary to create a different slightly more general target type list.

All repositories studied had very similar file structures, made up of a metadata page, sometimes a full metadata page, and when available, the actual resource itself, or in some cases resources (see Figure 7). All except the metadata page may or may not be present in the repository. Whether the link was to the metadata page/full metadata page or to the actual resource itself was noted. If the link was to the metadata page, we also noted if the resource was available within the repository and could potentially have been linked to.

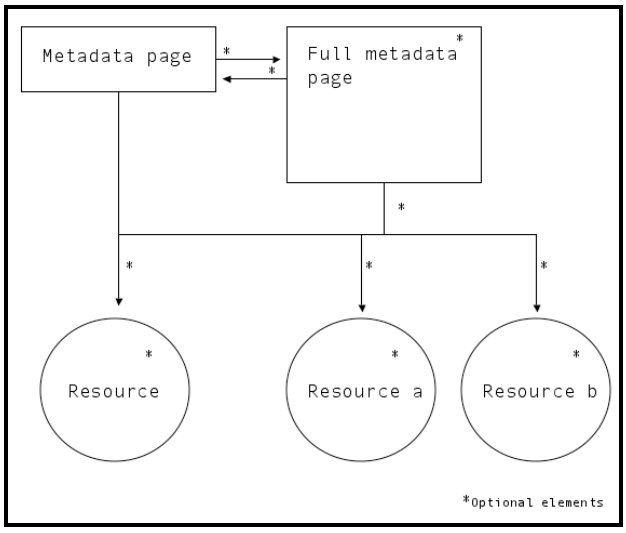


Figure 7- Repository file structure

Source page typology

All source pages were viewed and classified. As mentioned previously, classifying web pages is particularly difficult as there are still no standards. Some previous typologies exist (Almind and Ingwersen 1997; Haas and Grams 1998; Crowston and Williams 2000) and these were reviewed. Almind 1997 offered a classification system according to the function assigned by the author which was useful as a starting point but rather too general for this study. This was also true of Haas-Grams 1998, typology but which served as an excellent starting point. Crowston defined types of genres rather than genres, and looked at familiar genres on the web, new but accepted ones and ones that are apparently new. Some work has also focused on understanding link motivation creation (Cronin, Snyder et al. 1998; Haas and Grams 1998; Kim 2000; Thelwall 2003b; Wilkinson, Harries et al. 2003; Bar-Ilan 2005). However, the scope of the present study did not specifically seek to determine why the link had been created but rather to what item in the repository was linked to. An interesting area for further study would be to look further at link motivation. However, there is enough evidence from link motivation studies to show that links are an important source of information on the relationship between the

linking and the linked resources (Kim 2000; Wilkinson, Harries et al. 2003) and that certain web conventions related to the creation of links are emerging (Bar-Ilan 2005).

Defining source page typology is an important area of research in itself, consequently it was considered outside the scope of this study to look at this in too much depth. It suffices to say that the inherent difficulties in web page classification are an important indication of the continuing evolution of web publishing. The implications of this for scholarly publishing and communication are addressed further in the *Results and Discussion* section of this thesis.

An initial set of one hundred links was classified. These were then classified again a couple of weeks later and the two were compared. Discrepancies were resolved and the classification system was fine-tuned. Ideally, all the links for a study of this type should be classified by two people but this was not possible due to time and resource constraints. As this study is more concerned with target repository pages than classifying source pages, it was deemed sufficient. Target page and source page typologies are presented in the Results section.

Overview of repository registers and growth of repositories

The previous sections have described how the online questionnaire, and interview and link data from the case studies, all fit in together to address the research questions. The main scoping framework for data collection is the universe of institutional repositories, thus it was considered important to include in this methods chapter an overview of the main repository listing tools used for this research.

With the increase in repository numbers worldwide several registers have been created in an attempt to track their growth, the types of materials they collect, the number of items they contain and the overall growth worldwide. These repository registers have been particularly useful for producing advocacy data on the growth in the number of repositories online and as a

tool to help people find repositories by different criterion, such as type, geographic region, name and so forth. These directories were used frequently in the present research. It is therefore important to describe and analyse the methods used to put together these repositories in order to adequately interpret the data they provide and which is used in this research.

The following repository registers will be reviewed:

- **ROAR** (Registry of Open Access Repositories)
- **OpenDOAR** (Directory of Open Access Repositories
- **OAIster:** a union catalogue of digital resources in OAI compliant repositories

ROAR (Registry of Open Access Repositories)

Run by the University of Southampton in the UK, this service was launched approximately at the same time as EPrints software in 2000. Initially its objective was to serve as a list of repositories using EPrint software. This has since been expanded to include all repositories, independent of the type of software employed. The service is also a tool to help promote Open Access through pre and post print literature[32].

Registering a repository and collected info

Repository managers are encouraged to register their repository using an online form. Information is requested on the name and type of the repository, software used, country (if applicable), open access mandated, if full text is available and a short description field. The OAI-PMH base URL is solicited to track the growth of the repository. According to the information available on the service, repositories that are duplicated, inappropriate, non-functional or considered webspam are not entered into the directory. The main form of checking is using OAI-PMH base URL in order to harvest the metadata records. If this works appropriately then the growth of the items within the repository is also registered.

[32] For more information see the FAQ section for ROAR found at: http://trac.eprints.org/projects/iar/wiki/FAQ

Overall repository growth worldwide

ROAR appears to be the longest living repository directory. The first data available are from 1989 and since 2001 we have seen steady growth in the number of repositories worldwide, reaching over 1200 repositories by 2008. See Figure 8.

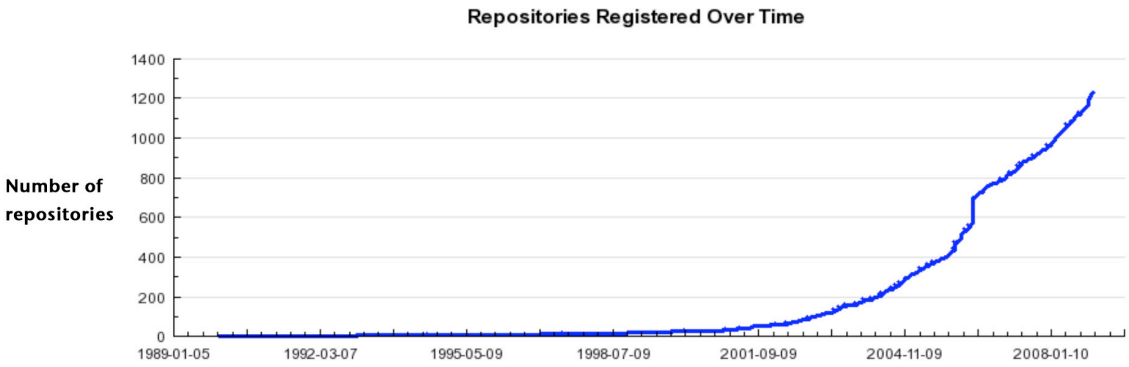


Figure 8 - Overall repository growth worldwide

Repository types and content

The directory registers repositories by different types. The pre-defined options are:

- Research Institutional or Departmental
- E-journal/Publication
- Other
- Database/A&I index
- Research Cross-Institutional
- Demonstration
- E-Theses
- Learning and teaching objects

From the ROAR webpage it is not clear how this classification system was determined. determined. Figure 9 shows the growth of different types of repositories over time. As we can see, Research Institutional or Departmental is easily the fastest growing type of repository.

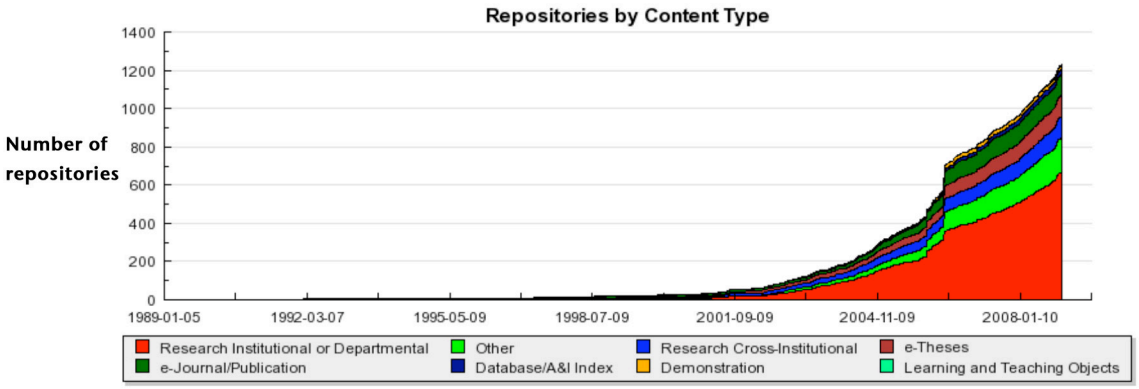


Figure 9 - Repositories by content type

There are currently 669 repositories registered as Research Institutional or Departmental[33]. Figure 10 shows this growth in more detail. What is particularly interesting is the steady and rapid growth of Research Institutional and Departmental repositories registered in ROAR from 2006 onwards.

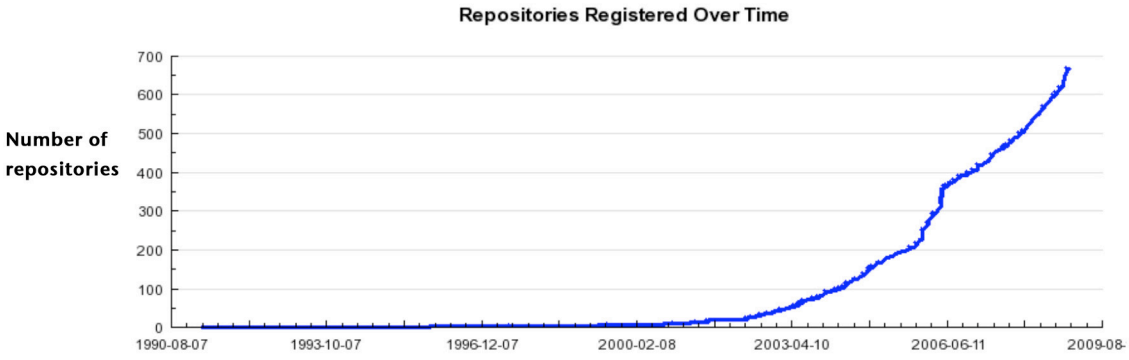


Figure 10 - Overall institutional and departmental repository growth worldwide

ROAR considers the repository type equivalent to the content type. The search box options for 'Content type' are the repository type options described previously (Research Institutional or Departmental, E-journal/Publication, Other, Database/A&I index, Research Cross-Institutional,

[33] Data taken from ROAR on the 15th December 2008

Demonstration, E-Theses, Learning and teaching objects). There appears to be an assumption that the repository type will define the content type.

Additional features

The ROAR collects and shows growth over time for individual repositories. For example, Figure 11 plots the growth of the UCL repository.

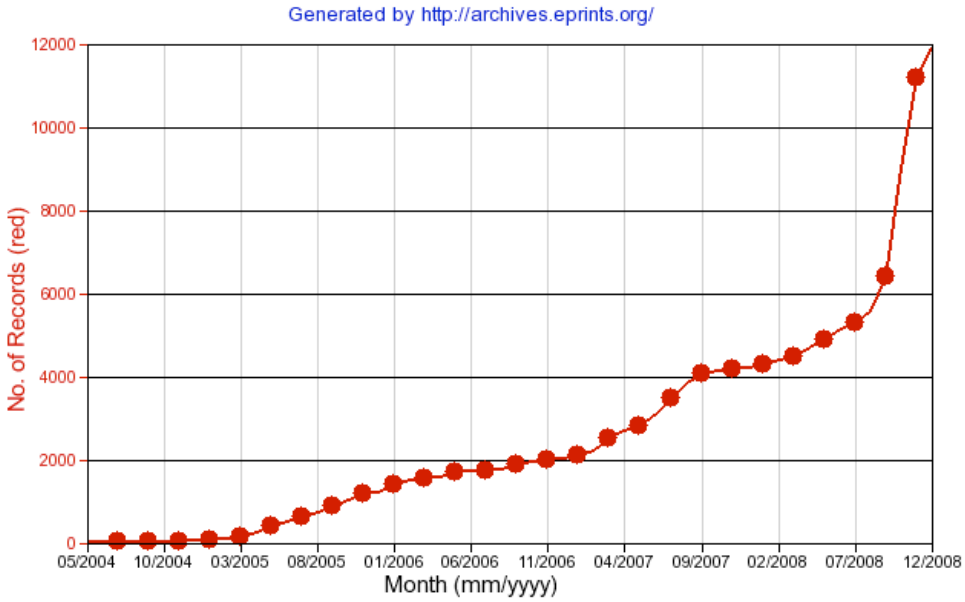


Figure 11 - Growth of UCL repository

Additionally, an interactive graph allows users to zoom into the data and look more closely at what was deposited on any particular day. An important caveat is that ROAR does not distinguish between metadata-only records and full text items. It is therefore not easy to appreciate if the growth in records is linked to a similar growth in deposited materials.

ROAR also allows searches by country, system software, content type and name. This provides valuable information on the selection criteria of repositories. This information can also be mashed with other applications, such as Google maps and Google Earth for graphic representation of the growth of repositories worldwide over time and geographically linked.

OpenDOAR

OpenDOAR was created in 2005 as a joint project between the University of Nottingham and University of Lund to register and categorize academic Open Access repositories. One of its mains aims was to act as a comprehensive and authoritative repository directory listing. It is currently maintained by SHERPA.

Registering a repository and collected information

Repositories are registered and categorized by the OpenDOAR team with the aim to provide a quality-assured listing that catalogues and describes the repositories it registers. This entails that each repository is viewed and evaluated manually. A description for each repository is written, the information about the repository is verified, the content viewed and so forth. Another main driver is to offer tools, such as search, filter, analyse and query, in order to facilitate user research.

OpenDOAR currently registers only repositories that are Open Access. This means that sites with any form of access control, such as passwords, or that contain only metadata records are not included in the directory listing. However, an important caveat is that OpenDOAR will register a repository if at least some of its content is OA.

The following are common reasons for not listing as stated on their web page[34]:

- Site is repeatedly inaccessible
- Site is an eJournal (OpenAccess or otherwise)[35]
- Site contains no Open Access materials
- Site contains metadata (bibliographic) references only or solely links to external sites
- Site is actually a library catalogue or collection of locally accessible e-books

[34] About OpennDOAR http://www.opendoar.org/about.html. Accessed 15th January 2009.

[35] A sister project DOAJ (Directory of Open Access Journals) run by the University of Lund aims to register Open Access journals. For more information see http://www.doaj.org/

- Site requires login to access any material (gated access) - even if freely offered
- Site is a proprietary database or journal that requires a subscription to access

OpenDOAR registers the following information about each repository:

- Name
- Description
- OAI base URL
- Type
- Software
- Size
- Subject
- Content type
- Languages
- Policies

Overall repository growth and repository types

OpenDOAR classifies repositories as one of four types: Aggregating, Disciplinary, Institutional and Governmental. OpenDOAR currently registers 1040[36] institutional repositories representing about 80% of the total database. Figure 12 shows the growth of IRs in the OpenDOAR repository. As with ROAR, there is a surge in the growth of IRs in mid-2006. However, the OpenDOAR team explained that this is most likely due to a backlog of work than actual reflected growth. The webpage states that from 2007 the graph better represents actual growth.

[36] Data taken from OpenDOAR on the 15th December 2008

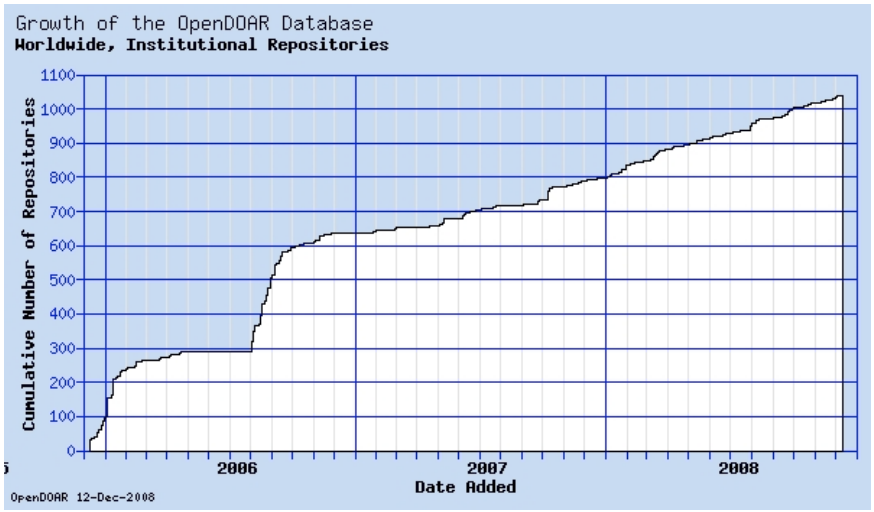


Figure 12 - Worldwide growth of Institutional Repositories

Repository content types

A main difference with ROAR is that OpenDOAR has made an attempt to classify the types of content within the registered repositories. The content types defined are:

- Articles
- Books
- Conferences
- Datasets
- Learning Objects
- Multimedia
- Patents
- References
- Software
- Special
- Theses
- Unpublished

This allows the user to view the number of repositories in OpenDOAR according to the types of content that they contain. For example, Figure 13 shows the content types for all the repositories in OpenDOAR.

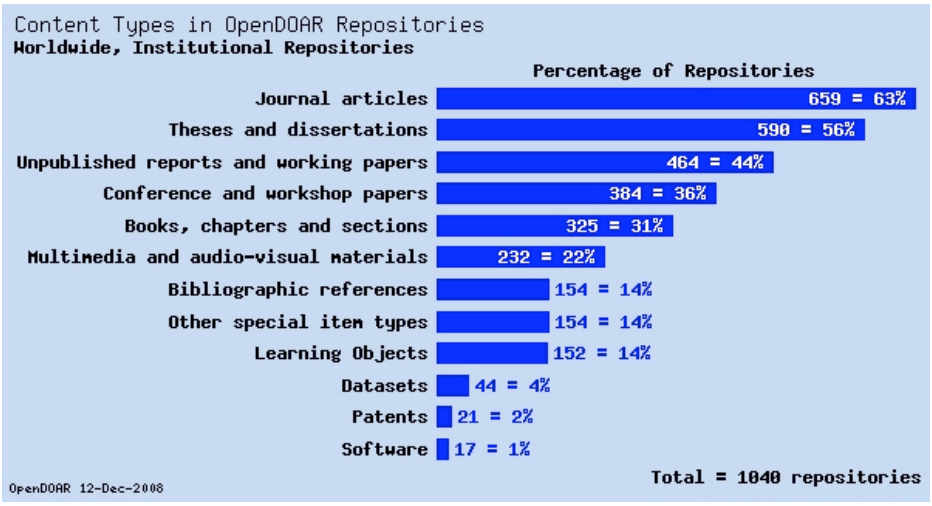


Figure 13 - Content types in OpenDOAR repositories

OpenDOAR however, does not provide information about the relative distribution of these different types of materials within the different repositories. The above graph indicates therefore that journal articles are the most popular type of materials to be accepted in their registered repositories but it does not indicate how many items are actually deposited. The tool indicates popularity for accepted content types but not for actual deposited types. It is however, still a useful first glimpse at repository content.

OAIster

This is a service offered by the University of Michigan since 2002. It is not exactly a repository directory but rather it is a 'harvester' of OAI metadata records. It currently has 1,051 contributors[37]. What is important to note is that these contributors are probably some type of repository either subject or institutional but in theory they can be any form of information service that uses OAI-PMH. For OAIster these are known as data providers. Therefore, a main difference with the other two directory services is that this one focuses on OAI but not on Open Access.

[37] Data collected on 15th December 2008

OpenDOAR and ROAR

These two repository listings are probably the best known and as mentioned previously, frequently used for advocacy. However, no systematic review of how these repositories are constructed was found which is an important consideration when these are to be used as research tools for mapping the repository landscape.

ROAR is the oldest registry and covers a quite substantial period of repository development. A large amount of data is available and it is possible to look at growth and changes over a long period of time. Additionally, the information and the graphs provided are dynamic and the information is therefore current. One of the main problems with ROAR is that it requires registration. This means that it only registers directories that have solicited insertion. The list may not be sufficiently popular or even meaningful enough for a lot of repository managers to be registered. Another important issue is that the quality assurance of the information on the repository is more dependent on technology than on human input. ROAR verifies that the OAI interface works but does not check up on the information provided. So for example, if a repository manager (or whoever registers) classifies a repository as an e-thesis repository then this is not verified. This may be particularly problematic for repositories in languages other than English or simply repositories that do not fit into the particular classification scheme. Figures on growth should be viewed with caution as they are indicators of the number of records within the repositories but not necessarily of access to full text.

The OpenDOAR registry offers more qualitative information and detailed information about each repository than ROAR. It is very clear on its inclusion and exclusion policies and in this sense, offers authoritative and trustworthy data. The downside is, of course, that the information is not necessarily up-to-date or dynamic. As we have seen repository technology and objectives

are in flux and it could be that a repository now accepts content types that it did not when the repository was first registered. This information is updated infrequently.

Table 10 shows a comparison of ROAR and OpenDOAR. This is an adaptation of tables presented by Peter Millington (Millington 2008).

OpenDOAR	ROAR
Search by keyword	Search by keyword
Filter by: Repository type, Software, Country, Language, Content type, Subject area	Filter by: Repository type, Software, Country
Analytical statistics	Content growth statistics
API, Policies Tool	Celestial - Harvesting analytics database
Repositories only	Repositories Some open access journals
Must have some open access full texts	Allows metadata-only and gated access items
Suggestions & Proactive discovery	Self-registration only
Manual validation	Taken on trust + some automated validation

Table 10 - Comparison ROAR and OpenDOAR

Summary

This chapter describes the research design and the different methods and procedures employed for data collection and analysis. The literature review as part of the methodology of this study aided in the reiterative design of the research questions and helped identify the key variables. This was followed by an online survey for repository managers in order to gain a better understanding of repository manager's attitudes towards non-formal electronic resources, their approaches to repository collection and depositing and their perception of the use and function of the repository. An email invitation to repository managers to reply to the questionnaire was sent out to selected email distribution lists. The response rate for the survey was calculated using the total number of worldwide repositories. The survey was also used to recruit candidates for a cross case study of repositories. Seven case study repositories were selected

from the UK and interviews were carried out with their repository managers. The aim of the interviews was to collect more detailed data about repository development, content ingestion work flows, depositing behaviour, content typology and resource usage monitoring. Log analysis and link analysis are discussed as methods for measuring usage. A qualitative link analysis study was done on all seven case study repositories.

The chapter describes how the methodology was applied and the difficulties encountered in particular with defining target and source pages. The relationship between the different data sets is discussed and how these fit together to address the aims and objectives of this study using a qualitative socio informatic approach. Additionally, this chapter also describes and analyses three of the most popular current repository listings: OpenDOAR, ROAR and OAIster. The following chapter presents the findings and discusses the results.

Chapter 4- RESULTS AND DISCUSSION

This chapter presents and discusses the results from the online questionnaire and the cross-case study of repositories where data was collected using interviews and link analysis. The initial section explains the rationale behind the presentation of the results and the overall structure. The remaining chapter is divided into seven parts. The first *–Demographics of the sample*[38] - presents the characteristics of the sampled repositories both from the online questionnaire responses and the seven case studies, including age, number of items and level of development. This information is contextual and its main purpose is to provide a general picture in order to better understand the results from the survey and the case studies. This section also discusses the representative value of the sample from the online questionnaire and how this was calculated.

This is followed by the presentation of the results grouped under six headings. *–Typology of electronic resources*- presents data about the different types of materials accepted within repositories and the methods for selection. *Distribution of electronic resources* shows results on the frequency distribution of different types of materials within the repository. The third section *– Depositors and workflow processes*- shows results on the levels of depositing activity and management of materials. *Repository objectives and drivers*- deals with the reasons behind the creation of repositories and the various purposes they may serve. The following section -*Usage and visibility of electronic resources*- presents the findings on usage statistics and other types of usage data from the survey and questionnaire data. In order to gauge visibility, this is followed by results from the link analysis study detailing a typology of all source pages that link to items within the repository. A typology of target pages is also presented as results. The final section

[38] Although demographics refer to the study of the characteristics of human population, the term has been employed here to refer to characteristics of the repository population.

Repository overview and general remarks presents data on general issues related to repositories and their future.

Presentation and structure

One of the issues that must be addressed with a research design that uses several data-gathering techniques is how to report and interpret the results. Traditionally, writing for quantitative approaches has relied on more conventional and predetermined formats, whilst for qualitative research the writing of results has been more varied and diverse (Punch 2005:260). However, the paradigm debate has also led to a rethinking of research writing and in particular the importance of the choices made when deciding how to present and structure results. A key component of research is not only the process of describing and analysing data but also the selection of the way in which the information is presented and structured.

Presentation of the results

Initially it was considered that the results from the online survey could be presented followed by the data from the individual case studies. In this way the survey data would offer a general, quantitative overview of repositories and serve as a benchmark of the current repository landscape worldwide. Subsequently each IR case study would be presented with the results from the interviews and from the link analysis in order to build a more focused and detailed picture about the different repositories.

However, this methods-driven way of presenting the results by focusing on survey, interview and link data did not seem the most appropriate manner and hindered rather than aided answering the research objectives and questions. So although it was thought that the more broad and quantitative survey data could not be combined effectively with the more in depth and qualitative interview data, this was not the case. One of the main reasons for this is that both the survey and the interviews focus on gathering data from repository managers and cover similar

subjects such as content types, repository objectives, depositors and use of usage statistics. By presenting the results by method the similar topics covered by both the survey and the interview were artificially separated and limited the analysis. Additionally qualitative data was gathered from the survey through several open-ended questions and a significant amount of data was actually collected in this manner. For example, the final question requesting general views or additional remarks about the types of electronic resources deposited in repositories and the groups who deposited was answered by 54 respondents providing a rich qualitative data source.

During data analysis of the open-ended questions from the survey and the data from the interviews using grounded theory the same codes and topics came up repeatedly. These mapped out quite easily to the variables that were defined in the Methodology section and it seemed appropriate to use these as categories as a means to finding a way to structure the results. Additionally, the link analysis data corresponded to several of the variables. In this way each one of the different data collection methods addresses several variables as shown in Table 11.

	Research variables	Questionnaire	Interviews	Link analysis
1	**Typology** of er*	√	√	√
2	**Distribution** of er	√		
3	**Collection** policies	√	√	
4	**Work flow** processes	√	√	
5	**Usage** of er	√	√	√
6	**Visibility** of er		√	√
7	Repository **objectives** and drivers	√	√	
8	**Attitudes** towards non-formal er	√	√	

* electronic resources

Table 11 - Research variables and data collection methods

Structure of the results

The research variables were classified and transformed into six headings under which to present the results. These broad headings also reflected the principle issues addressed in the form of research questions and the objectives. The headings are: typology of electronic resources,

distribution of electronic resources, depositors and work flow processes, repository drivers and objectives, usage and visibility and general remarks. The survey and the interview results are presented grouped underneath these headings. In order to interweave the results appropriately and avoid confusion, the data source is always indicated. In general the survey results are presented first and the interview data, with the appropriate case study number, is used to exemplify, clarify or broaden a particular result.

The link analysis data is treated slightly different from the interview and survey data. From Table 11 we can see that the link analysis addresses three data variables: typology, usage and visibility. Initially it was considered that the target page typology developed from the link analysis could be presented as part of the results underneath the typology heading. However, this fragmented not only the link data but also the actual analysis of all the links. It seemed more effective to present the link data and analysis all together within the visibility and usage section. Additionally the link analysis was designed specifically to address the question of visibility and usage and although it does contribute towards our understanding of typologies of electronic resources, its main contribution is within the visibility and usage arena.

Demographics of the samples

In order to contextualize the results from the survey and the case studies, this section provides contextual information about the repositories from which data was collected either from repository managers or through links. This information is important to place the subsequent results within context in order to inform the analysis and interpretation of the results. Additionally for the survey the representative value of the sample is addressed.

Repository demographics from survey

Country distribution

Responses were received from a total of 31 countries (see Table 12). The most represented countries were the United States (20% of responses), closely followed by the United Kingdom (19.4%). This was followed by Germany (5.3%) and the Netherlands (4.7%). Prior to the targeted email invitations, only one country, Brazil, with 2 repositories was registered from the Latin American region. Following these emails, Mexico went up to 6.5% with 11 responses and Brazil to 4.1%. Regionally speaking there was a strong European representation, with 52.5% of the repositories, closely followed by North America[39] with 23% of the total responses (see Figure 14).

Country	Frequency	Percent
USA	35	20.6
UK	33	19.4
Mexico	11	6.5
Germany	9	5.3
Netherlands	8	4.7
Brazil	7	4.1
Italy	7	4.1
France	5	2.9
Australia	5	2.9
Switzerland	5	2.9
India	5	2.9
South Africa	4	2.4
Japan	4	2.4
Canada	4	2.4
Poland	4	2.4
Spain	3	1.8
Norway	3	1.8
Denmark	2	1.2
New Zealand	2	1.2
Portugal	2	1.2
Sweden	2	1.2

[39] For this study North America is comprised of Canada and the United States. Mexico is placed within the Central American region. This is done in order to compare with the OpenDOAR data that uses this classification system.

Paraguay	1	.6
Finland	1	.6
Mongolia	1	.6
Croatia	1	.6
Ukraine	1	.6
Belgium	1	.6
Austria	1	.6
Greece	1	.6
Malaysia	1	.6
Colombia	1	.6
Total	170	100.0

Table 12 – Survey responses per country

Representative value of sample

In order to assess the representative value of this sample the relative percentages of country representation were compared to the OpenDOAR registry[40]. The number of repositories per country in OpenDOAR was compared to the number of r epositories per country in the survey. Figure 14 shows the data by geographical region and Table 13 shows by individual country.

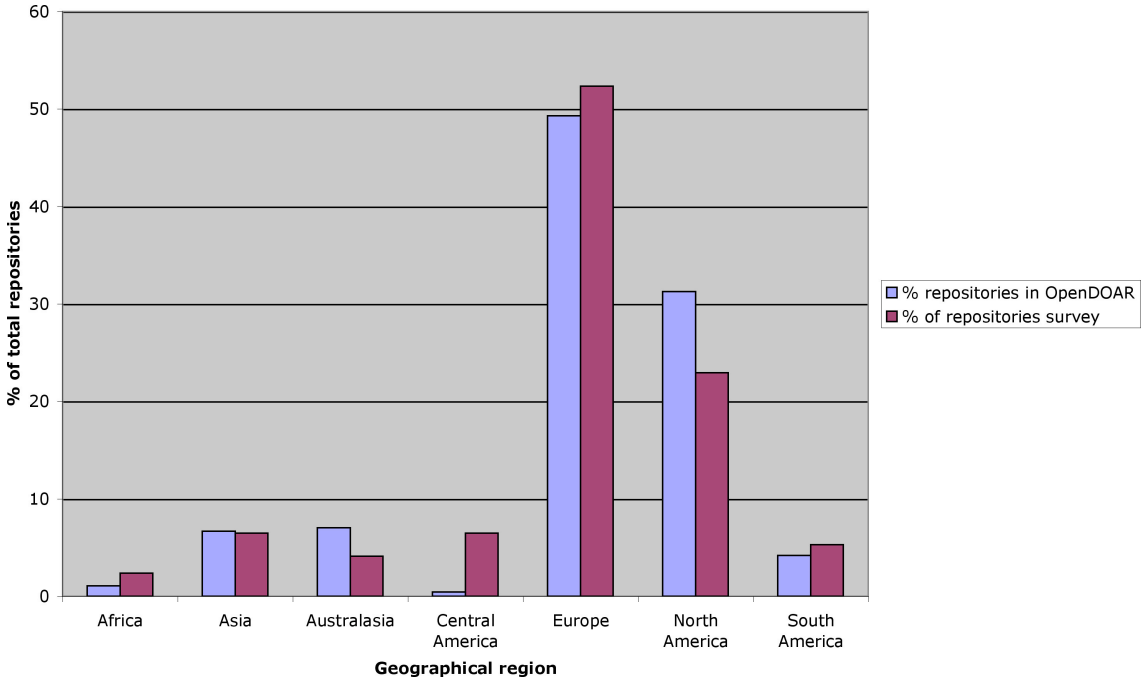


Figure 14 - Number of repositories per region for survey and OpenDOAR

[40] OpenDOAR data is from the 18th of September 2007.

Of the countries with a large number of repositories Germany, Australia and the United States appear to be under represented in the survey sample, whilst the UK is overrepresented. This is most likely due to the fact that three of the email discussion lists were focused mainly on UK subscribers (JISC-repositories, JISC Cetis Metadata and UKCORR[41]). Of the countries with a smaller number of repositories Belgium and Sweden had a lower response rate. In the preliminary analysis India was greatly under represented (0 responses) and the targeted email invitation helped make the figure representative. However, for Latin America the targeted email created some imbalance by over representing Brazil and in particular Mexico (0.32% in OpenDOAR and 6.47% in survey data, see Table 13). Poland and South Africa are also slightly overrepresented. However, it is clear from the survey data that there are more repositories in Latin America than OpenDOAR registered and this could help to explain the figures. In addition, OpenDOAR has selection criteria, as described in the Methodology chapter for registering a repository, which the survey did not have. The remaining 21 countries appear to be well represented in relation to the OpenDOAR registry.

Country	# repositories OpenDOAR	% of total (927)	# repositories survey	% of total (170)	Difference between OpenDOAR and survey
Australia	52	5.61	5	2.94	2.67
Austria	5	0.54	1	0.59	-0.05
Belgium	16	1.73	1	0.59	1.14
Brazil	24	2.59	7	4.12	-1.53
Canada	31	3.34	4	2.35	0.99
Colombia	2	0.22	1	0.59	-0.37
Croatia	3	0.32	1	0.59	-0.27
Denmark	6	0.65	2	1.18	-0.53
Finland	8	0.86	1	0.59	0.27
France	34	3.67	5	2.94	0.73
Germany	115	12.41	9	5.29	7.12
Greece	4	0.43	1	0.59	-0.16
India	20	2.16	5	2.94	-0.78
Italy	28	3.02	7	4.12	-1.1

[41] See Annex 2 for a list of all email lists with descriptions.

Japan	32	3.45	4	2.35	1.1
Malaysia	1	0.11	1	0.59	-0.48
Mexico	3	0.32	11	6.47	-6.15
Mongolia	0	0	1	0.59	-0.59
Netherlands	44	4.75	8	4.71	0.04
New Zealand	13	1.4	2	1.18	0.22
Norway	6	0.65	3	1.76	-1.11
Paraguay	0	0	1	0.59	-0.59
Poland	9	0.97	4	2.35	-1.38
Portugal	4	0.43	2	1.18	-0.75
South Africa	9	0.97	4	2.35	-1.38
Spain	17	1.83	3	1.76	0.07
Sweden	31	3.34	2	1.18	2.16
Switzerland	6	0.65	5	2.94	-2.29
UK	104	11.22	33	19.41	-8.19
Ukraine	2	0.22	1	0.59	-0.37
USA	259	27.94	35	20.59	7.35
TOTAL	927	95.8	170	100	

Table 13 – Repositories per country for OpenDOAR and survey

To the best of my knowledge this is the first systematic worldwide survey of repositories. Previous surveys have focused on a smaller number of countries and aggregated the data at a national level (Westrienen van and Lynch 2005) or have been conducted as country surveys (Lynch and Lippincott 2005; McDowell 2007; Rieh, Markey et al. 2007).

Survey and case study combined characteristics

Repository stage development

Managers were asked to classify the development of their repository according to three different stages: Prototype, Recently launched/Initial stage and Fully operational repository. No definition for these different stages was provided but the question presented no difficulties during the pilot testing or during the survey, as the categories are fairly self-explanatory. They are also quite similar to other classifications used in other studies, for example both US studies (Lynch and Lippincott 2005; Rieh, Markey et al. 2007).

115 of the repositories, 67.6% were described as *Fully operational* with only 5.9% registered as prototypes. 38% were considered to be recently launched (Figure 15). From the case studies five repositories were considered fully operational, one recently launched and one was categorized as recently launched although it had been fully operational for some time but had undergone extensive redevelopment. If we take into consideration that repositories are a fairly recent development, from 2002 onwards, then in the space of five years a relatively large percentage of the repositories consider their repository to be fully implemented. However, as we shall see in the interview data, there is a difference between having a fully functional *technically speaking* repository and having a repository that is fulfilling its storage and dissemination functions.

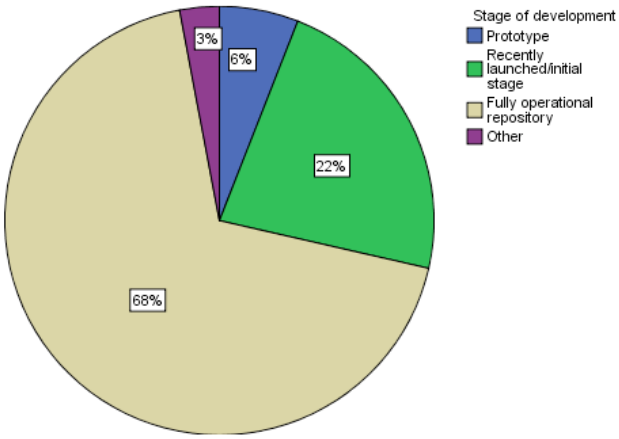


Figure 15 - Repositories by stage of development

The results can be compared to the MIRACLE census of IRs (Rieh, Markey et al. 2007) that censed repositories by stage of development in the US. Their sampled population also includes respondents with no IR and over 50% of the respondents did not have and did not plan to develop an IR. Of the remaining respondents 20% are only planning to develop an IR, 15% are actively planning and pilot-testing and only 10% have an implemented IR. The MIRACLE survey results seem to indicate that penetration or IRs in the US higher education systems has been limited and that few institutions actually have an IR.

One main difference between both studies is that the online survey for this research was directed only at repository managers, so respondents would have to have an implemented IR whilst the MIRACLE report surveyed all university library directors. Additionally these results show the problems of defining the 'institution' in the institutional repository. The 13 nations survey (Westrienen van and Lynch 2005) reported difficulties when collecting their data due to the fact that there exists a variety of approaches to defining what constitutes a university. For example, the US survey (Lynch and Lippincott 2005) considers 261 universities whilst MIRACLE is looking for IRs in over 2,000 institutions. Based on this, the Lynch and Lippincott survey found that 40% of universities had an IR compared to MIRACLE's finding of 10.8%. In other countries, the 13 nations survey found that countries such as the Netherlands, Norway, Germany reported that 100% of their universities have an IR.

In the case of the data collected from the online survey, one of the main limitations is that repository managers were not asked to define the type of repository that they were running and in retrospect this would have been a useful question to ask. This means that the data collected could be about an IR or another type, such as a subject repository. However, respondents named the organization hosting the repository and in the vast majority of cases it was easy to identify the institution as a university. Additionally from the information provided by OpenDOAR and ROAR it is known that most repositories worldwide are institutional repositories. Therefore, although the survey data may contain a few repositories that are not institutional, this is not enough to alter the validity of the results.

Repository age

In the survey we found no relationship between the age and stage of development of repositories. We would have expected to find that repositories required a certain amount of time to be considered fully operational but we found fully operational ones in the less than a year old

category. The MIRACLE survey found that of the IRs that were operational, over 50% said that their IR had been operational up to 1 year, 27% from 1-2 years, 4.2% from 2-3 years months and 16.6% more than 3 years. Our data shows that 20% of the repositories were less than a year old, 24% 1-2 years, 20% 2-3 years, and over 20% were more than 3 years old, as shown in Table 14. The MIRACLE survey data is collected from US institutions whilst survey data is worldwide which could account for some of the differences. From the MIRACLE data there appears to be a recent boom in IR creation whilst the online survey data is fairly evenly spread over all categories.

Age of repository	% of repositories MIRACLE survey	% of repositories online survey
Less than 1 year	50%	20%
1-2 years	27%	24%
2-3 years	4,2%	20%
3 or more	16.6%	20%

Table 14 – Repository age MIRACLE and survey data

Number of items

We asked respondents to indicate the number of items held within the repository (see Figure 16). A little over 70% of the repositories reported less than 5000 items, with only 6 repositories above 50,000. However, 51 repositories (30%) reported between 1001 and 5000, with 23 repositories already between 10,000 and 50,000.

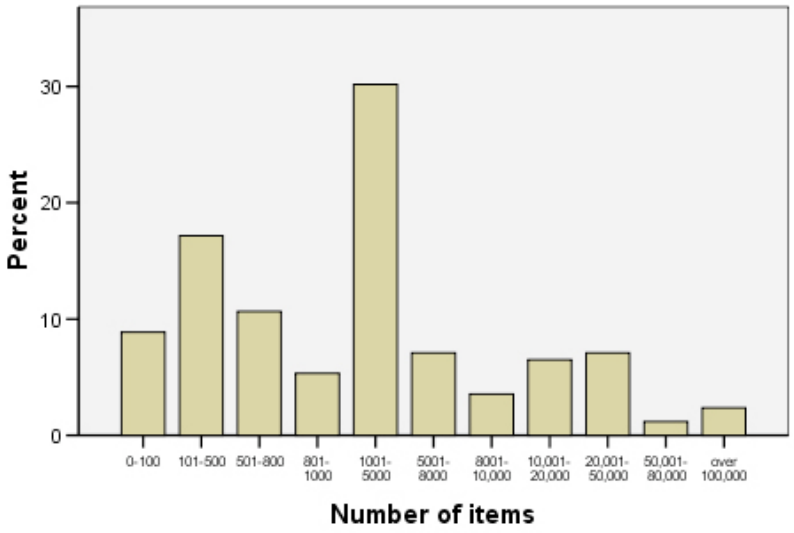


Figure 16 – Repositories by number of items

We found very similar results to the 13 nations survey (Westrienen van and Lynch 2005) that indicated the difficulties with counting number of items or objects within the repository. The 13 nations survey found that some repository managers considered metadata records only as items within a repository whilst others only counted an item if an object is attached to the metadata record. Some repositories may only contain bibliographic references fed in, so although the repository seems quite large, in terms of actual digital content it can contain little. If dissemination and access are key drivers for repository development there are key issues to be discussed when interpreting figures of for example, 14,000 metadata-only records compared to 1,000 records with full text attached. For our survey this was not clarified and it is not clear whether the numbers indicate metadata only or full text.

This problem of counting items did not come up in the survey data but it did during the interviews. At the beginning of the interview, managers were asked to confirm the number of items that they had indicated when responding the survey and almost all had some kind of caveat or comment to make about forms of counting.

> We have about 12,000 in our repository of which, I mean you can slice and dice this in lots of ways. *[CS2]*

One case study introduced an additional issue with counting items and these are items that have an object attached but are not accessible outside of the institution or sometimes even at all; these are the so-called dark archives.

> It depends on what question you are asking. How many Open Access, full-text items in the repository… it is about three and a half thousand. There is another one and a half thousand items which are restricted either within the University or more restricted than that. And then we have a number of metadata only records. *[CS5]*

One of the consequences of expanding the range of types of materials collected in a repository is that some of the materials are actually not for dissemination at all but rather just for management, storage and preservation. These types of materials may have severe restrictions on their dissemination due to a number of reasons, confidentiality being one of them. Examples of these are administrative documents or sensitive datasets. One case study had the extreme example of even having to hide the metadata from search engines due to data protection laws. They had digitized class photographs and stored them in a dark archive. The metadata was available for harvesting and contained the names of the students in the photographs.

> We actually had to add a patch to DSpace so that we can have dark items that are completely dark so that you can't even see the metadata. Because the names of these people are in there and we can't give those out because that is personal information. We have to be very careful with that. *[CS1]*

Although it can be argued that size alone is not always an indicator of success respondents were aware of the importance of reaching a critical mass but unsure of how to define it. As put by one of the interviewees:

> You need the numbers to make the repository look… I still don't quite know what the critical mass is but it needs to demonstrably have stuff in it for people to start taking it seriously. *[CS6]*

Respondents suggested several ways of interpreting the significance of the number of items within a repository. One repository for example, that was not too keen on counting the number of items as a measure of success, looked at the number of papers by their staff published by a key publisher in their field and compared this to the number of pre or postprints from this publisher in their IR. In this way they found that every single paper published with the key publisher was deposited in the IR. This indicated to them 100% coverage. This example of course, is only applicable to formal publishing but demonstrates another way of determining success and which is directly dependent on repository drivers. In order to properly interpret repository item numbers we must know not only what repository managers are actually counting but also these figures must be viewed in light of the repository objectives.

Case study demographics

The case study repositories are all from the UK. The reasons for this are described in more detail in the Methodology section. However, the selection offers a variety of characteristics in terms of type, age and number of items. Table 15 shows a breakdown of repositories by age, type, number of items and repository software.

Case study	Type	Software	Age (years)	Num of items
CS1	Fully operational	DSpace	4 - 5	Over 100,000
CS2	Fully operational	Eprints	More than 5	10,000 – 20,000
CS3	Fully operational	Eprints	2 - 3	101 – 500
CS4	Recently launched	Fedora	Less than 1	101 – 500
CS5	Fully operational	Eprints	3 – 4	1001 – 5000
CS6	Other (launch/fully operational)	Eprints	2 – 3	1001 – 5000
CS7	Recently launched	Open repositories	1 – 2	101 – 500

Table 15- Type, age, ·num of items and repository software by case study

Typology of electronic resources

Although the literature review indicated that in general repositories are set up to manage and disseminate 'research output' in most cases it is not clear how repositories decided upon the types of materials that they accept and in some cases it was not even clear what types of materials they accepted at all.

Types of materials allowed for deposit in repository

In the questionnaire, respondents were asked what types of materials were allowed for deposit within their repository. As mentioned in the Methodology section a list of material types taken from the OpenDOAR registry service was used. The different types are listed in Figure 17.

Books and book chapters	Preprints
Conference proceedings	References/bibliographies
Workshop papers	Software
Datasets and databases	Theses and dissertations
Learning objects	Reports
Audio-visual materials and multimedia	Working papers
Patents	Images, maps, diagrams
Postprints	Administrative documents

Figure 17 - Content types for survey

Three of these categories did not appear in the actual online survey due to a composition problem in the HTML when using Survey Monkey. Therefore, there is no data available for Working papers; Images, maps, diagrams and Administrative documents. It was not intended that this list be definitive and so this omission did not have major implications on the usefulness of the data collected. One possible limitation is that the online survey data is no longer comparable to the data from OpenDOAR, which would have been desirable. However, since undertaking this survey the categories in the OpenDOAR directory have been altered, so even if the online survey had collected data from all the categories, the subsequent modifications would have still made comparisons difficult.

Originally respondents were asked to indicate not only the type of material that they accepted but also whether this was the peer-reviewed version or not. However, once this data was collected this differentiation proved troublesome and rather than providing insight it confused matters. For example, the peer-reviewed status of bibliographic references is generally irrelevant as these are not usually subject to peer review. Similarly it could be assumed that postprints are peer reviewed and preprints are not. In this sense peer-review was not a very useful additional piece of information and because of this it was decided that it was better to merge the information about peer-review and the results are presented together.

Respondents could select as many item types as they wanted and Figure 18 shows the percentage of repositories accepting certain resource types.

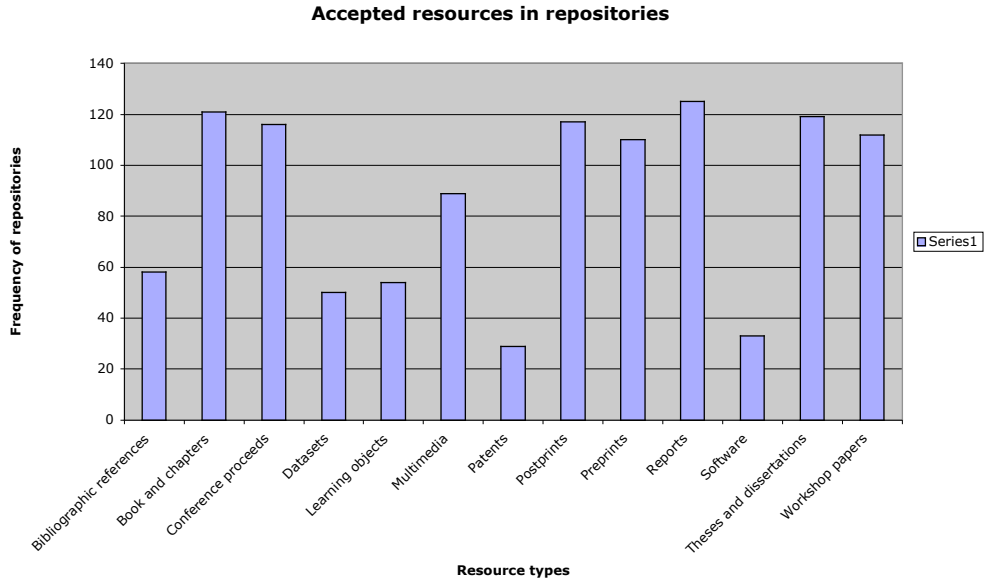


Figure 18 - Types of materials accepted by repositories (n=170)

The most common type of resources accepted within repositories were reports, very closely followed by books and book chapters and theses and dissertations all above 70%. Between 60 and 69% of repositories would accept postprints, conference proceedings, workshop papers and preprints. It is quite clear that the major categories are paper based digital equivalents that are well embodied in the scholarly communication process. More digital only objects, such as software, datasets and learning objects are lower in the scale.

It is not a surprising find that the IRs more commonly handle textual materials as these are important in the scholarly communication process and both academics and librarians are familiar with these types of formats and use them often. An interesting result is that reports were the most popular type, even more so than preprints and postprints. It would have been reasonable to expect that almost all repositories would accept journal articles. However, if postprints and preprints are merged into one category, e-prints, then the number of repositories accepting this type of material rises to 76% making it the most popular type of resource. This is not surprising considering that the origins of IRs are from e-prints, however it seems that accepting either postprints or preprints but not both is a practice by a number of repositories. This is an interesting find considering that originally IRs were designed to archive preprints that eventually would be replaced if possible with the postprint.

Over 50% of respondents indicated that they accepted multimedia, a generic umbrella term that usually covers a wide array of materials such as video and audio. An issue with these particular results is that the question was framed in such a way that respondents could only select from the options presented to them. This naturally limits the types of materials that could be selected as the list presented was not exhaustive. However, it was useful in serving as an initial benchmark and respondents were then asked to please note other types of electronic resources accepted within their repository and not included in the list.

Additional materials accepted for deposit

Over 100 of the 170 respondents answered this question. The breadth of materials is proof of the wide variety of electronic resources available and the results further indicate the need for work in the development of digital content typology. The answers were analysed, grouped, repetitions eliminated and similar types assimilated in order to produce the list in Figure 19 showing the materials in alphabetical order.

- Administrative documents
- Architectural drawings
- Blog posts
- Book reviews
- Committee papers
- Conference power point presentations
- Confidential directorate documents
- CVs
- Departmental newsletters
- Digitized materials general
- Digitized materials-XML encoded texts
- Digital art
- Digital photographs
- Essays
- Exam papers
- Executive Orders and Proclamations
- Guides/manuals
- Institutional memory types of documents (i.e. historical photos)
- Lectures
- Magazines
- Maps
- Musical scores
- Newsletters
- News reports about research
- Photographs and photograph albums
- Poetry
- Projects/business plans
- Promotional flyers
- Psychological scales
- Raw and processed imagery obtained from satellite and aerial platforms
- Raw data generated in projects
- Records of court proceedings
- Research memo
- Research project proposals
- Results of observations Results of observations and simulations related to publications in peer-reviewed articles (Astronomy & Astrophysics)
- Scans of old books.
- Student's assignments
- State mandated public access materials
- Teaching material
- Tutorials Syllabus
- Speech
- Videos
- Web pages

Figure 19 - Materials allowed for deposit in repositories, not in original list

One respondent added:

> Later down the road we hope to add unpublished data and databases (…) images of faculty and student artwork, recordings of music faculty and students as well as symphony, video of university sponsored plays and from the film students and images of the buildings and institutional art that is displayed all over the campus. Other prospects are under debate but have not been formally discussed.

Some of these non-formal resources from the list share common characteristics and can be easily grouped together. For example, committee papers and confidential directorate documents are types of administrative documents; scans of old books, historical photos and so forth are digitized materials and others such as exam papers and student assignments could be within a teaching and learning category. However, the range is quite broad and some resources such as poetry, architectural drawings and news reports about research are more difficult to categorize. Some of these resources seem to be digital only materials with no print counterpart and could be

considered new digital resources. Examples of these are XML encoded texts, blog posts, digital art and digital photographs.

What is particularly interesting to note is how much of this can actually be classified as 'research output'. It seems that quite a lot of IRs are actually defining their materials types in the broadest sense possible and following Lynch's definition, using it for the "management and dissemination of digital materials created by institutions and its community members" (Lynch 2003:328), including but not limited to research output.

Using peer-review/published as a factor

Some survey respondents indicated that they accept all types of materials in principle but that they do take into consideration other factors in order for the material to be eligible for deposit. These factors are generally related to concepts of published and peer-review. In some cases repositories will only accepted materials that have been published and/or peer-reviewed in some form or another.

Using peer-review or published as considerations for accepted types of materials is a form of quality assurance for the resources within the repository. This is a surprising find considering that the origins of IRs are in preprints, articles that had been submitted for publication but not necessarily peer-reviewed or published yet. It seems that as IRs have developed, there is a growing need to find mechanisms that certify or validate the materials that they offer. If we return to the concepts of publishing examined in the literature review one key issue is ensuring quality (Tenopir and King 2001). Although IRs are taking on certain roles of publishers such as collecting and providing distribution and access, other roles such as peer-review and the validity of 'published' are still apparently left to traditional actors.

> The primary criterion for inclusion is "published research material". Resources need not have been peer-reviewed but they must have been "published".

These considerations, published and/or peer-reviewed, also directly affect the range of types of materials that will be accepted within a repository. For example, it is highly unlikely that datasets, learning objects, interview transcripts or images will pass the criterion of peer-reviewed or published. Indeed it is quite clear that repository managers of these types of repositories are thinking primarily and maybe even exclusively of formal digital publications. Interestingly enough these types of IRs do not fall into any of the IR trends identified in the literature review for they do not accept preprints nor do they cater to a broad range of digital materials. It could be argued that a new type of IR may be developing that provides records and access to the institution's formal research output exclusively.

We found evidence, in particular from one of the case studies that had a peer-review only policy in place, that in practice this was actually very difficult to implement. The boundaries of the concept of published are unclear, in particular with digital grey literature and non-formal resources. Focusing mainly on journal articles and book chapters, this case study repository's collection policy is that the item must be published and peer-reviewed. This means that they do not take preprints:

> So we don't take preprints, that is one of the things that we definitely was clarified quite early on. We only take material after the peer review process. *[CS7]*

However, when asked specifically about working papers, reports and conference proceedings they indicated that this was not as clear-cut:

> Again it has been fudged a bit…Initially there was very strong guidance that was in place when I first started working on the repository was that it must be peer

> reviewed. But in practice we were already taking material that wasn't peer reviewed because we took the papers from [centre named deleted] and they are actually not peer reviewed". *[CS7]*

To solve this problem the managers of the IR eventually broadened their scope and instead of just accepting peer-reviewed materials they changed it to 'published' materials by which they meant that it must have an ISSN or ISBN and should already be in the public domain with the university's name attached to it. In this particular case the IR is deliberately not taking on a publisher's role and refuses to include (and therefore certify) any type of material that has not been published or made accessible by other members of the university previously.

In other cases, IRs will accept both peer-reviewed and non-peer reviewed material but will indicate, presumably in the metadata, whether a particular object has been peer reviewed. For example, one survey respondent mentioned:

> Peer review material is always marked up so you can see status of the publication, also pre and post prints

Judging from the case study experience it would be fair to assume that repositories that take this approach probably grapple with the same issues relating to defining whether a document has been peer reviewed or not. However, as mentioned in the literature review validation and certification are unresolved issues even in the print world. Kling has argued that scholars have sophisticated methods that rely on various processes and markers to indicate the trustworthiness of a particular resource (Kling McKim 1999). For the time being IRs are using indicators, such as peer review and published as these are familiar and well known but it could be that in the future, in particular for non-formal resources, IRs begin to develop new types of quality and validity indicators.

Dealing with non-formal electronic resources

There is an important difference between what repositories would accept in theory and what they actually have accepted for deposit within the repository. Although the list of non-formal resources is quite extensive this does not necessarily mean that repositories actually have this material. As put by one survey respondent:

> Please note that I had confusion with the question above - our repository WOULD ACCEPT all of the types of materials listed; we don't necessarily HAVE all those types of materials.

Repository managers raised several issues regarding the difficulties involved in managing non-formal resources and the different approaches taken towards solving these issues.

Capacity to handle non-formal resources

Although IRs would in theory accept diverse materials one of the main difficulties was the repositories ability to handle a wide variety of material types.

> Others from this list will be accepted, but we are not quite set up to receive them yet.

What is not particularly clear from this quote is whether they are still not set up technically speaking or if there are other issues such as metadata, preservation, certification and workflow processes affect a repository's capacity to handle non-formal electronic resources. As mentioned in the literature review it has been argued that the main issues are non-technical although there a few more recent studies (Emly 2007; Salo 2008; Shreeves and Cragin 2008) have described technical limitations with the software, in particular for managing non-formal resources.

We found that repository managers had several approaches to handling non-formal electronic resources and these can be grouped into three different approaches: creating a separate repository, dealing with new types of items on an *ad hoc* basis and a wait and see approach.

Creating a separate repository

One approach was to consider that certain types of materials did not actually belong in the institutional repository and should be better deposited elsewhere. This can be either another repository within the same institution or some kind of external, possibly subject type repository.

> I strongly believe that very specialist material, such as scientific data or learning objects should be stored in specialist repositories (…) rather than in one general all-purpose repository.

> We will accept anything that is a research output and we feel is not best placed elsewhere e.g. specialist repository also in the institution (-name deleted-) or external (large scale data).

Having more than one repository in an institution is not unusual. In at least two case studies for example, a separate repository had been created within the institution specifically for learning objects. In one case it was not that the IR could not handle learning objects but rather because it was not considered the best place for them to be. The IR focus was on dissemination and access to electronic resources and they considered that learning objects needed to be stored but not disseminated. As put by the manager "and the feeling politically is that academics don't want them visible" [CS3]. The rationale behind this is that learning objects can be considered more like raw material from which a lot more can be extracted, like creating a course, whilst research is an output and is therefore for dissemination. The other repository with a separate learning repository also considered that their repository was only for research and did not cover teaching and learning.

Another repository case study was also creating a separate repository specifically for multimedia materials, mainly because they would be using software that would be better at handling this type of materials than your "bog standard e-print repository" [CS6]. Additionally the manager considered that there were very different drivers behind the e-print repository that is for access and dissemination of research output (mainly journal articles), and the multimedia repository that is more focused on curation and preservation of these digital materials.

> The other thing (…) is that there are separate drivers for the e-prints repository which are not necessarily related in any way in the mind set of the people who are likely to give us the money to develop it to the principles of digital curation. I mean we are talking about maximizing research impact through visibility and positioning [the institution] in any kind of metrical analysis to the best possible standard. Which is absolutely and entirely different to the sort of drivers that are making [the multimedia repository] develop. So there is some logical separation but I agree that in the longer term physically there must be a bit more unity." *[CS6]*

In another case a repository manager would suggest that an item type be deposited in an alternative location, outside the institution, when they felt that the institutional repository was not adequate. One example is a repository that would accept datasets only if they did not consider that they would be better off elsewhere, specifically dynamic datasets, which they deposited in a national database. They felt that this service was better suited to handling this particular type of item, both in experience and in the software that they employed.

Additionally another manager questioned the concept of datasets in itself.

> "The other thing is that everybody talks about data as this sort of… What is data? I mean yes to us it is datasets but it is images, its… We have had something from the English department they want to use us, which is sort of an annotated thing, that's data. So data can be anything and

> we are going to get it, it is just a question of feeling our way with all that difficult stuff" *[CS4]*

An example of this is a case study repository that has the collected letters of a writer from the nineteenth century classified under database/dataset category [CS6].

Ad-hoc basis

The second approach, used by two repositories, is to look at new types of materials on a case-by-case approach and decide whether it should be accepted.

> If we get a request for a deposit which is slightly unusual or doesn't fit our usual parameters, we will talk about it as a group and decide whether it sets a precedent and what issues there might be about taking or not taking it. *[CS5].*

A particular example is a podcast of an interview with an academic done by a local news channel discussing his research. The repository committee discussed and although it was not a research output as such, it did help to conceptualize and enhance the visibility of the university's work and was deemed appropriate for the repository.

Another repository manager felt that it was the academics themselves who introduced the need for new item types within the repository. Originally although the repository accepted a variety of items, the academics would tend to refer to their repository as containing only peer-reviewed items as a means of assurance and quality. On one occasion an academic deposited an audio/video file and after this event other academics followed suit, altering the peer-review only rule:

> And then one of the professors put in an avi file which was a recording of a broadcast news item which had mentioned their work (…) and of course as soon as that happened and other people started to do that as well. But it was the professors themselves that broke rank. If I had suggested

> that a year beforehand they would of said oh no that is not refereed. But as soon as one of them appeared on the national news they couldn't put it in the repository fast enough because it was a way of enhancing their profile, boasting about their research. *[CS2].*

Textual orientated repositories are aware that in the near future issues about these types of new materials will have to be addressed. Non-textual material, such as video, audio, images and datasets are going to play an increasing role within their repository. A few of the repositories were participating in projects and studies designed to discover how to deal with non-formal materials, including cultural materials and datasets.

> We are looking to expand the range of the types of materials that the repository will hold. Again with this philosophy of we want to represent the full scope of the research output of the institution, that is one way of doing that, ensuring that you can cater for the breadth of material. *[CS5]*

In this sense, some repositories are well aware that catering for only digital formal publications will tend to leave out digital materials created by subjects or disciplines within the university that do not rely so heavily on journal articles or even books as a means of communicating research. As members of the steering group the School of Art and Design had voiced this aspect of their research output.

> The School of Art and Design were there and they were going to provide the first non-textual things like photographs, sound and video. *[CS7]*

However, these types of materials cannot be handled in the same way at selection, ingest, metadata, copyright, access and preservation level as other items. They are looking to work on test materials to devise methods and metadata.

Wait and see approach

In some repositories, although there was an awareness of the issue, it was clear that currently it was not a priority. Most repositories were focusing on capturing the journal articles. However, the repository managers tend to be mindful of this type of item.

> There are some departments which are obvious sources for this material [videos, images, audios], which we haven't yet tackled and we are aware of that. *[CS6]*

> So we are going to have to deal with anything that comes our way (…) we had to say for now, well we can't really do anything other than basic textual stuff. *[CS4]*

> There is clearly interest and awareness but we know it is a new order of challenges as well. *[CS6]*

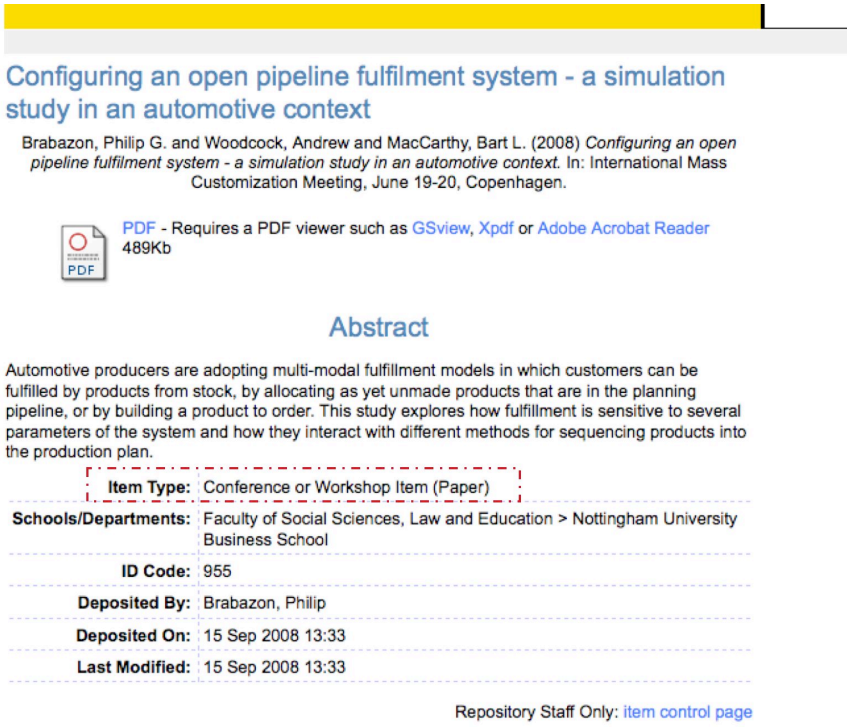
Additionally, most repositories seemed to be aware that for the future it would be more about integrating not only a wide range of types of materials but also a number of different repositories, both institutionally and with other subject based repositories.

> These are the things we need to look at in the future and further develop in terms of interoperability and making sure everything joins up effectively (…) So given that we are never going to have one repository solution, we have to decide when it is best to create a separate repository for a purpose. *[CS5]*

Almost all repositories were looking towards several repository solutions within their institution and using a federated system approach to integrating them all under one umbrella. Issues about the relationship between institutional and subject based repositories also came up, although not in the context of non-textual types but in relation to journal articles.

Definition of typology lists

In the repository case studies it was not clear from any of the repository websites how the different types of materials that they accepted had been agreed upon. All case study repositories include in the metadata for the digital object an 'item type' (or similar) field in which the genre of the object is stated. Figure 20 shows a screenshot of a repository item metadata page including the item type field. Either through self or mediated deposit, this field is usually defined by selecting a genre from a predefined list.



Configuring an open pipeline fulfilment system - a simulation study in an automotive context

Brabazon, Philip G. and Woodcock, Andrew and MacCarthy, Bart L. (2008) *Configuring an open pipeline fulfilment system - a simulation study in an automotive context.* In: International Mass Customization Meeting, June 19-20, Copenhagen.

PDF - Requires a PDF viewer such as GSview, Xpdf or Adobe Acrobat Reader
489Kb

Abstract

Automotive producers are adopting multi-modal fulfillment models in which customers can be fulfilled by products from stock, by allocating as yet unmade products that are in the planning pipeline, or by building a product to order. This study explores how fulfillment is sensitive to several parameters of the system and how they interact with different methods for sequencing products into the production plan.

Item Type:	Conference or Workshop Item (Paper)
Schools/Departments:	Faculty of Social Sciences, Law and Education > Nottingham University Business School
ID Code:	955
Deposited By:	Brabazon, Philip
Deposited On:	15 Sep 2008 13:33
Last Modified:	15 Sep 2008 13:33

Repository Staff Only: item control page

Figure 20 - Screenshot of metadata with item type

It was not clear for any of the repositories how this list had been decided upon and what considerations were taken into account. As this had also not been particularly clear from the survey either, during the interviews with repository managers they were asked how these genre types had been defined. Interestingly discrepancies were found with regard to the types of

materials that the seven case study repositories stated that they accepted for deposit, actually accepted in practice and planned to accept in the near future.

As one would expect repositories tended to start with formal electronic publications. The content types were developed from what was likely and known: "But they are what you would expect from what is being produced by researchers" [CS4]. These lists tended to be restricted to digital items with print counterparts such as journal article, book, book chapters, conference proceedings and so forth but they are aware that this list will grow. There is "a lot of digitized stuff that needs a home" [CS4]. Managers tended to be aware that in the future this list would increase or be modified.

It is particularly interesting to note here that OpenDOAR modified its genre type list during the period between the online survey and the writing up stage[42]. In some cases the category name was shortened with no significant change to the concept. For example, Audio-visual and multimedia was shortened to just multimedia and theses and dissertations was shortened to theses. Some of the more notable changes are the merging of postprints and preprints categories into one articles category. A special category was added and working papers and report categories disappeared and a new unpublished category created. Administrative documents and images, maps and diagrams also disappeared. The old and new category types are presented in Table 16.

New	Old
Articles	Preprints Postprints
Books	Books and book chapters
Conferences	Conference proceedings
Datasets	Datasets and databases
Learning objects	Learning objects
Multimedia	Audio-visual materials and multimedia

[42] Online survey concluded in September 2007. The changes in OpenDOAR were noted in September 2008 during the writing up period.

Patents	Patents
References	References/bibliographies
Software	Software
Special	
Theses	Theses and dissertations
Unpublished	Reports Working papers
	Administrative documents
	Images, maps, diagrams

Table 16 - New and old content types from OpenDOAR

In one case study repository there was a strong discrepancy between the types of materials that they stated or listed as accepted in the repository and what actually happens in practice. For example, within the browse section a variety of item types were available. However, when questioned directly about one document type 'performance' the manager replied "I have no idea what that is" [CS3]. Moreover, when asked how this list was developed and what criterion was considered for it, it appears to have been developed very early on in an unsystematic fashion and there are no plans for reviewing it:

> It was something that was decided on day zero. I get the feeling it was because that it was what the software could cope with at the time. *[CS3]*

So, although an extensive and rich document type list exists, there is no documented rationale or policy behind it.

> I get the feeling it was one of those things that we just said, we will take those things (…) if you can find any actual policies you are doing well. They are just sort of semi-acknowledged truths. It is a terrible thing to say. *[CS3]*

The list included: artifact, show/exhibition, composition, performance, image, video, audio, dataset and experiment, and yet when questioned directly about specific deposits it emerged that

multimedia and datasets are still under discussion for acceptance within the repository, despite these item types listed as accepted within the repository. Their first focus is still on getting materials, namely e-prints, into the repository and then they will look into these other types of materials. In particular datasets, as there has been interest from some departments.

It was clear to the manager that these inconsistencies between the types of materials that they indicated they would accept and what they could actual accept for deposit in practice reflected badly on the seriousness of the repository.

> "- What can go in there?- ,
> -Oh well anything… We don't really take data… -I know it says data-
> And so well you have left that and it makes you seem like a very unprofessional service" *[CS3]*

The experience of designing the survey and using the OpenDOAR content types and the ensuing changes in the lists with no apparent explanation on why or how this was done, resonates with repository manager results.

Distribution of electronic resources

Using the same original typology list from OpenDOAR[43], respondents were asked to indicate out of the types of materials that can be deposited in their repository, what they considered to be the three least and most frequent types of resources. Respondents were not obliged to select any and the maximum was three.

Figure 21 shows the number of responses as a percentage for the most and least frequent resource type within repositories.

[43] As there was no HTML error in this section when designing the survey, data for images, maps and diagrams; administrative documents and working papers is included.

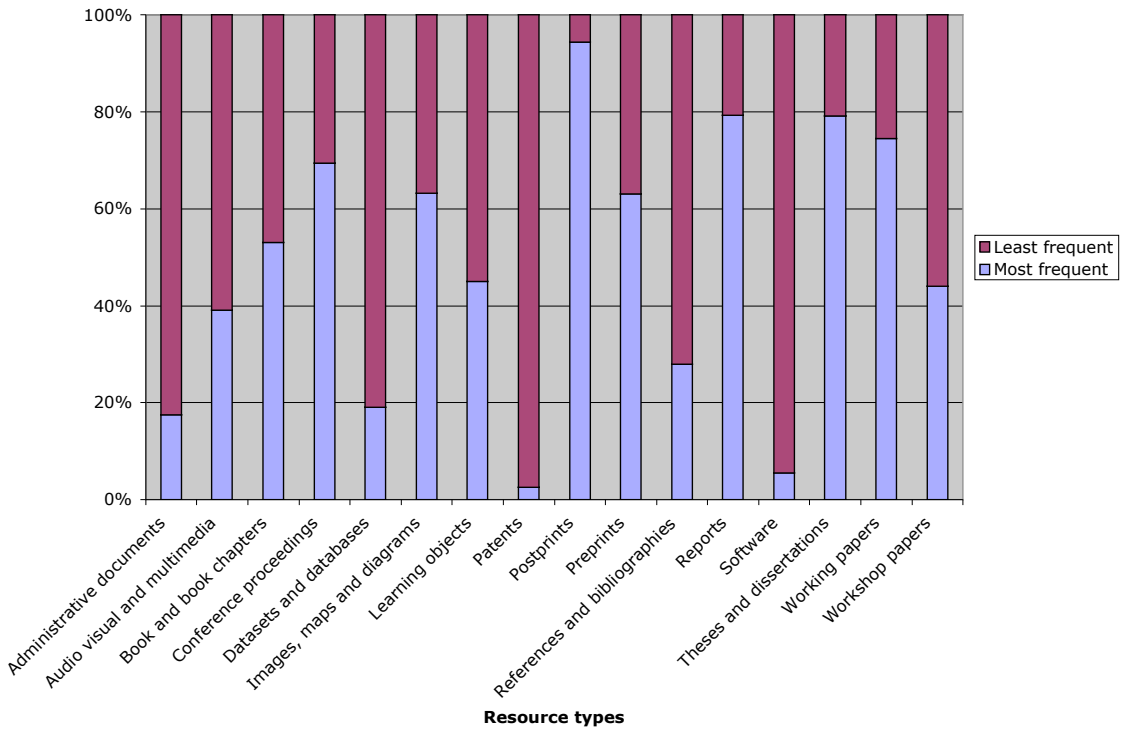


Figure 21 - Most and least frequent content types in repositories

The most frequent types of materials contained within repositories were clearly postprints followed by reports and theses and dissertations. The least frequent types are patents, datasets and databases, software and administrative documents.

However, other materials, such as book and book chapters are reported by almost the same number of respondents (34 to 30) as both the least and most frequent type of material. There is a similar situation with learning objects and workshop papers. Preprints, one of the common formats under discussion in the context of repositories, did not figure as one of the most frequent types and is below reports, working papers and books and book chapters in terms of frequency.

These results were compared to the data gathered from three other repository surveys reviewed in chapter two: the Ware study of 45 repositories (Ware 2004a), the 13 nations survey (Westrienen van and Lynch 2005) and the MIRACLE survey (Rieh, Markey et al. 2007) that

focused on the US. The results are presented in Table 17. The 13 nations survey's data is presented by country and countries with no data are not included. The MIRACLE survey results are divided into data from repositories in the planning and pilot testing (PPT) stage and the ones with a public implementation of an IR system (IMP).

In order to make the data comparable in the studies where the data was presented as a percentage this was modified into a frequency. So for example if a survey indicated 70% e-prints, 20% ETDs and 10% reports, this was changed to 1, 2 and 3 respectively, with 1 indicating that document type as the most frequent and 3 as the least frequent.

Types	**Ware**	**13 Nations**						**MIRACLE**		**Survey**
		Australia	France	Norway	Sweden	Netherlands	UK	IMP	PPT	
E-prints/ Articles	2	2	1	2	2	2	1	3	2*	1**
ETDs	3		2				2	1	1	3
Book & Book chapter				1	1	1				
Learning objects										
Working papers	1							2		4
Reports										2
Primary data		1						4		
Others										

Table 17 - Frequency of content types among various surveys

* Preprints

** Postprints

In Ware's study the largest percentage (58) of all materials is grouped under the category 'others ' that includes mainly grey literature but also some digital images. E-grey literature and non-formal resources have fuzzy boundaries and tend to be grouped all together under the same category. Australia reports that about 83% of its IR content is primary data which is a very high

percentage of non-formal resources. In the MIRACLE study for implemented repositories (IMP) and the online survey both working papers and reports make up a substantial part of IR content.

As expected, e-prints is a popular type in particular in the UK, France and worldwide online survey data. However, one important limitation of the 13 nation survey data is that no difference is made between preprints and post. In the online survey and in the MIRACLE survey this distinction is made although results are contradictory. In the latter, repositories in the pilot testing stage reported a larger percentage of preprints, whilst for the online survey, where almost 70% considered themselves fully implemented repositories, there was a larger percentage of postprints. This data could suggest that IRs, at least for journal articles, move from self-archiving preprint beginnings towards a more formal library-run repository that focuses on record keeping of formal publications.

However, interpreting this data is particularly difficult and the exercise itself particularly enlightening about the difficulties in trying to understand what types of materials are found within IRs. Two things in particular are clear from this data:

1.1. The data is extremely difficult to compare due to a number of reasons. One of the main problems is the different ways that materials are categorized. For example, three countries group books and ETDs together underneath the same heading. In order to understand the impact that repositories can be having on scholarly communication these types of groupings are not particularly useful as the role of books and the role of theses are different within the formal scholarly publishing and communication system. Additionally book and book chapters are sometimes grouped underneath the same category that again defocuses the data. As mentioned previously pre and postprints are not usually separated either. The scenario for non-formal resources is not very detailed, especially if we take into consideration the variety of types that are accepted within IRs.

1.2. When looking at frequency or percentages of materials within a repository it is difficult to know what it being counted. As mentioned with overall repository sizes when looking at numbers of different types again we do not know if these figures refer to metadata only records or include the actual digital object. So although a repository may report 90% of journal articles it could be that these are all metadata only with no full text. The 13 nations survey in particular is highly aggregated data that was collected quite unsystematically and as mentioned by the authors offers an initial overview but is not necessarily reliable or specific data (see Westrienen and Lynch 2005).

Depositors and workflow processes

In the online survey, data was collected on depositors of items and during the interviews this information was complemented with further details about how materials were managed from ingest to final acceptance in the repository.

Repository depositors

In the survey respondents were asked who decides on what types of materials can be deposited in the repository. The options: myself, library, special committee and computing services, were offered as well as a free text other option. No definition for these options was included as from the piloting of the survey they were deemed fairly self-explanatory. A special committee was the most frequent response, followed by the library and myself (see Table 18). It appears from this that the content part of repositories is identified as an area that deserves particular focus, involving the formation of a special committee. However, there are not great differences between the frequencies of the three most popular groups. The least popular option was computing services, which would indicate that although their participation is limited to the technical side of things and the decisions about content are relegated to other groups.

Group	Number of responses	Percentage
Special committee	73	43%
Library	66	39%
Myself	52	31%
Computing services	4	2%
Don't know	1	1%

Table 18 - Decision-makers types of materials

Respondents were also given the option to name other decision makers in relation to the content policies of their repository. A large number of respondents indicated that deposits were defined by a community administrator, by each program or by faculty. One respondent indicated that the library set down broad guidelines but each local community decided on the particulars. From these findings it seems clear that content deposit in repositories is not a one size fits all solution and that many have devised a flexible approach by community.

> Each department sets collection policy, usually based on what faculty can declare as "academic work" in their annual reports. Engineering has a very different definition of "academic work" compared to the Fine Arts departments.

Research output of a university is therefore not necessarily defined at a global institutional level but rather each individual community will define what within their particular field could be considered a research output.

For some other repositories, content type is governed by entities outside the community of users. Specific responses were: state mandated, research body, university press and scientific editors. In some cases there were also editors defined for each section. It is quite likely that some of these repositories are subject rather than institutional and this could explain the differences in the bodies defining content.

Other repositories indicated an open approach by stating that it was the author/depositor who decided on the content type to be deposited. One particular interesting response remarks that although they would accept all content types they could only commit to preserving certain types of materials. This is one way of dealing with the problem of preservation and diversity of content types.

Respondents were then asked what user groups were allowed to deposit within their repositories. The options to select were: lecturers and researchers (LR), library staff (LS), administrative assistants (AA) and students (ST). These groups were not explicitly defined but are fairly recognizable to most university members. As would be expected the most common type of user groups allowed to deposit are LR and LS. The most infrequent user group is ST, as shown in Table 19.

User groups	Number of responses
Lecturers and researchers	117
Library staff	117
Administrative assistants	72
Students	50
Don't know	1

Table 19 - User groups allowed to deposit in repositories

The free text option presented a wide variety of answers that seems to indicate that a number of repositories have developed sophisticated user group authorization structure, depending on their different needs. Some repositories had a complete open deposit system in which anyone (presumably who registered) is allowed to deposit. These are probably subject-based repositories as opposed to institutional ones. Institutional repositories tended to have a more varied authorization structure, in which other user groups included general repository administrators or community administrators to deposit material. Some even mentioned that determining user groups authorized for deposit is organized by collections within the repository,

with each community being allowed to define for themselves their depositors. Other repositories are still building on a case-by-case basis, allowing individuals to contribute once their credentials had been checked.

Although students was mentioned as a user group within the option, many respondents used the free text option to specify that it was graduate students but not undergraduate students that were allowed to deposit, suggesting a marked difference between these two subgroups. This may be related to perceptions about types of material as well as quality control and level. Research assistants were also mentioned. For repositories that are used to publish journals, article authors and publishers were mentioned user groups. Finally, another user group was partner institutions indicating collaborative repositories.

Of the user groups that are authorized to deposit, respondents were asked about the level of depositing activity in two scenarios: one as authors of the deposited material and the other as depositors for material in general, whether as authors or on behalf of a third party. Respondents were asked to rate their activity in one of the following options: rarely or never active, occasionally active, frequently active or extremely active, as shown in Table 20. In the case that a particular user group was not authorized to deposit, this could be indicated by the respondent.

	Not authorized to deposit		Rarely or never active		Occasionally active		Frequently active		Extremely active	
	F	%	F	%	F	%	F	%	F	%
Lecturers and researchers	19	12	33	22	67	45	19	12	12	8
Library staff	13	9	46	35	36	28	18	13	18	13
Students	58	44	20	15	26	20	16	12	10	7
Administrative assistants	44	36	47	39	21	17	8	6	0	0

Table 20 - Level of deposit activity by user groups as authors

The results indicate very low percentages of user groups being extremely or frequently active, with the highest being library staff, 13%, stated as being frequently or extremely active. However, 45% of respondents considered lecturers and researchers to be occasionally active. This is a surprising result and it would be important to compare this with actual deposit rates in repositories. Much of the literature has frequently indicated poor academic participation (Duranceau 2008).

One of the user groups that was not contemplated when deciding on what options would be offered to respondents was repository staff. It was assumed that these would be working as part of the library staff but in the free text option several respondents indicated specifically that an important amount of depositing activity was done by repository staff. We can only presume that they meant as authors and not on behalf of a third party as this was what the question addressed. It would be fair to expect repository staff to be active in self-archiving their own work as an example to other members of the university- a practice what you preach approach. However, one case study interview mentioned that they deliberately tried not to include library materials.

> The initial items were no big surprise in that they were one or two things from the librarians, who obviously produce things. We actually took a decision at this stage that we were not going to take articles written by librarians really if we could avoid it. (…) there are a lot of repositories that I look at that have got a lot of items in them but they are all to do with the library and you begin to get this sort of feeling that within the institution it is the libraries repository, the library is running it, filling it with library things. *[CS3]*

It is important to point out that from the piloting of the survey these two questions, regarding the differences in depositing activity by user groups depending on whether it was as authors or as third party depositors, caused some confusion. It is clear by the open text responses that this

was not completely ironed out. It seems to take a while for people to grasp the difference between self-archiving and third party depositors. This is an important point as many repositories are hoping to depend more on self-archiving and less on staff or zealous academics depositing on behalf of others.

Respondents were then asked to rate the depositing activity of the same user groups but as depositors in general, both for themselves and/or on behalf of a third party. See Table 21.

	Not authorized to deposit		Rarely or never active		Occasionally active		Frequently active		Extremely active	
	F	%	F	%	F	%	F	%	F	%
Lecturers and researchers	16	11	28	19	62	44	21	14	13	9
Library staff	11	7	16	11	28	20	37	26	48	34
Students	54	42	22	17	26	20	14	11	9	7
Administrative assistants	27	21	29	23	25	20	35	28	8	6

Table 21 - Level of deposit activity by user groups in general

The level of participation, in particular for library staff goes up significantly, with 34% and 26% indicated as being extremely and frequently active respectively. This points towards library staff undertaking a great deal of depositing work on behalf of others which is the exact opposite of the self-archiving ideal. This also shows that Thomas and McDonald's assumption that the depositor of an item will be the author is not necessarily true (Thomas and McDonald 2007) and future studies on depositing behaviour should avoid this supposition. Depositing patterns are much more complex. The lecturers and researchers' deposit activity goes up slightly but not significantly. The level of student activity goes down, suggesting that their activity may be limited to their participation as authors of theses and not much else, as a large percent, a bit over 40% are not allowed to deposit in general. Once again administrative assistants are not well represented.

Respondents could also rate the activity of other user groups. Publishers, repository administrators, journal editors, archive staff and special committees were deemed to be extremely active. This was followed by: PhD students, partner institutions, management and general public as occasionally active. So although the library staff is by far the most active, there is a wide and varied group of people depositing in repositories that as a whole seem to make up for a substantial amount of the depositing activity of a repository.

The survey addressed the issue on amount of depositing activity by different user groups but did not look at this in terms of disciplinary differences. However, one survey respondent mentioned that depositing activity varies in terms of discipline, with the humanities and social sciences being more likely to deposit than the hard sciences.

> Contribution to institutional repositories varies widely depending on the field of research. Humanities and social science researchers are much more likely to deposit items in a repository than are researchers in the hard sciences.

So far there has been no research into disciplinary deposit rates. One hypothesis could be that hard sciences are more likely to deposit more formal electronic publications (such as e-prints), whilst humanities and social sciences may have higher deposit rates for other types of materials, such as audio files, annotated texts and others. However, for IRs that are not accepting non-formal types of resources then the arts and humanities and possibly social science deposit rates would be lower, simply because their normal publishing and communication vehicles are not catered for. This issue must be addressed if IRs are to capture and provide a home for all university research output.

Item ingest and work flows

In order for an item together with its metadata to be ingested within a repository, a number of processes must occur. Repository managers were asked about the workflow process involved. During the interviews the details and nuances of how this actually works shed a great deal of light on the complicated issues that occur behind the scenes and are in stark opposition to the commonly upheld notion that depositing in a repository is simply an act of an academic depositing an item with a few keystrokes.

One of the key issues that considerably affect workflow process is the introduction of quality assurance for items within the repository. In almost all the case studies repository items and their metadata were checked by a cataloguer (or similar title) before appearing in the repository. Quality assurance tended to involve checking metadata and copyright clearance for journal articles. Almost all repositories reported a backlog in the items for quality assurance checking.

> We almost don't want everyone to start sending us everything actually. We don't have the infrastructure in place to properly deal with it if they did. *[CS7].*

It seems fair to assume that there can be a considerable time lag between the moment an academic deposits (or submits an item in the case of mediated deposit) until it actually appears in the repository. In the case of non-textual materials that are examined on an *ad hoc* basis or for those repositories that are still unsure of how to manage them, this can probably cause even further delays. Several managers reported this time lag as a major disincentive for depositing.

Quality assurance of metadata

One repository initially did not have any type of quality assurance and they believed that the systems worked more efficiently when they did not. The rationale was that the metadata was created mainly for search engines, namely Google, and if there were slight errors this was not

such a big issue as the most important aspect was increasing access and dissemination, an objective that was achieved in this way.

> Things just got deposited. If you are not going to use them for administrative purposes, if they are just for researchers, then you put it in, Google reads it, you type in a query, you find it. Even if you make mistakes about the citation, what does it matter? *[CS2]*

However, as more demands were made on the repository for added on services such as for example, RAE or citation counts, they have had to move towards tighter quality assurance of the metadata. This factor came up frequently in the interviews as using the repository for other functions meant that the metadata had to be of much better quality.

> We never used to do anything about QA because it conflicted with the goals of the repository.(…) But if you are actually going to return this, you need accurate metadata for the RAE or anything else like that then you have to put the QA processes in place. *[CS2]*

Metadata that has been inputted by academic staff could be considered adequate for searching and retrieving functions, that is to say the dissemination side of a repository but is not reliable for other repository functions such as the RAE returns.

Back end batch deposits

One repository does not regularly accept front end deposits but works mainly with back end batch deposits and must be mentioned separately as it is a particular case. In this instance the main problem is that new mechanisms and workflow processes are created specifically on each occasion depending on the type and order of the new items. The repository team is planning to involve academics before creating the data, at the planning and project application stage, in order to streamline the ingest process at a later stage for the repository. Additionally they are

also looking at several options for front-end deposit. The current front-end deposit interface is cumbersome for multiple deposits and taking into consideration the nature of their item types, mainly videos, images, audios and so forth, single deposit interfaces are not a viable solution.

For non-formal resources the design of the interface of the most common software (namely EPrints and DSpace) is an important concern. These were designed for single uploads of preprints. Unless specific add-ons are designed (which requires technical skill) if for example, the English Department wants to upload five hundred digitized letters, these would have to be done one by one. This is not a practical or viable solution. As repositories move towards accepting new types of materials, that are not necessarily a single article, this is an important consideration.

Dedicated repository staff

One major development for most repositories in recent months has been increasing the number of staff dedicated to the repository. Up to this point most repositories had been run, usually on a part time basis, by just one person. Now most repositories have at least one full time staff member and some two or three. The most important tasks that can now be completed are advocacy and quality assurance. One repository manager feels that this will clear the way, once the day to day part of running a repository is more catered for, to begin to address more substantial and advanced issues.

> It will raise the priority of the repository and it may well bring up just these sort of questions that we are touching on now. *[CS3]*

Full time dedicated staff will also increase the amount of time dedicated to advocacy and strategic planning of the role of the repository within the university. Indeed the allocation of

funding for hiring full time staff in most cases was an important recognition of the repository as an important service provided.

> It changed from something that the university can just let tick along in the background to becoming something that you need support as a central service *[CS3]*

Additionally, when asked if the academics attitude towards the repository has changed, most repository managers mentioned a more general awareness, if not complete acceptability.

> The view people have of it has changed as well, initially from something unknown to now something that is known, half-trusted, half-mistrusted. *[CS1]*

However, it still appears that there is lack of general awareness of the exact function and benefits of a repository. The institutional repository in this sense is still, in most cases, on the fringe of university:

> [The repository is] on the edge of the university, doing something that is vaguely important, but that no one quite comprehends. *[CS3]*

Repository managers frequently mentioned that a large percentage of their time is spent on advocacy but it seems that IRs are still not central to the university. However, data gathered from the interviews indicate that the general perception is that this is a critical period and that things are set to change.

Repository objectives and drivers

In order to better understand why IRs were collecting certain types of materials and what they expected or wanted to achieve by doing this it, data was collected on the history and the main drivers behind their creation.

Repository origins

During the interviews repository managers were asked to describe the origins of the repository, why it had been created and how the initial items were collected and deposited. The history and early motivation for creating the repository could be key to its development as well as to understanding the types of items that can and will be accepted for deposit. In particular the first items deposited can set a trend for the direction in which the repository grows.

Of the seven case studies, five stemmed specifically from repository development projects and the initial items were acquired in a variety of ways. One repository [CS3] initially focused on the Sciences as it was considered that these subjects have more of an inclination towards sharing. They also followed up personal contacts that supported Open Access and who had shown interest in sharing their materials. They did find however, that certain groups such as the Physics community that already had an established subject repository (arXiv) were not keen on depositing in the IR.

For the rest of the university a great deal of time has been invested in advocacy and talking to different members promoting the repository. Despite all sustained advocacy work the repository manager still feels that most people are not aware of the IR or about OA.

> I think here it really has been true advocacy, really how to advocate not just the repository but the advantages and the whole OA thing because even after five years the vast majority of people do not seem to appear to know too much about it. I would love to say that my job now with the repository is so easy things just flow in but they don't. *[CS3]*

However, the repository manager felt that things could be at a turning point. There were two reasons for this. One was the fact that they had been able to hire dedicated staff as a result of a funding scheme from the university. The second was the they were looking into using the

repository to do RAE returns and this according to the repository manager would certainly boost the number of entries in the repository as well as making people more aware of its existence.

Another case study [CS6] took what they considered the "*well trodden path*" and targeted individuals on a one to one basis hoping to get some momentum. The repository was set up quite specifically for promoting open access and the main focus of collection were journal articles. It was also an early repository, developed at a time when it was considered that self-deposit could work without too much difficulty:

> We would demonstrate that the repository was useful [to the institution] and would inexorably lead to a mass adoption by academics who would throw their hats in the air and rush to give us all their papers. *[CS6]*

This however, had not been the case at all and the manager reported a very low deposit rate. However, in retrospect the manager realizes that if this would have happened they would not have had the capacity to deal with everyone depositing their research.

> We have always been constrained by resources in that we couldn't really mobilize anybody to go out and do [advocacy] systematically and because if the floodgates did open then we wouldn't have been able to deliver, you know, to put the stuff within the repository either. *[CS6]*

There are however, several recent events that lead to this repository manager thinking that things are about to change.

> The two very significant trigger events that have helped us to move it even further forward into the research flow have been firstly the research funders deciding that OA was a good thing. And then what has really tipped the balance is the move to metric based research and the realization that visibility of research is very important and that you can

> actually get this from the repository in a way that you can't from the conventional publications database you, that administrative database of record. *[CS6]*

Now that the repository has been in place for a few years and a full time member of staff has been assigned, the repository manager feels that the time is right for the repository really to really begin to grow in terms of items deposited.

Another case repository [CS5] was developed as proof of concept of moving from a successful existing departmental repository towards an institutional repository accepting deposit of materials from the whole university. They specifically targeted six different subject areas and systematically worked from there. Although not limited to e-prints, the departmental repository was quite text-based whilst the institutional repository sought from the start to accept different item types. When the repository began it was thought that academics would require assistance and an assisted mediated deposit scheme was put in place. Perhaps fortunately there was not a mass uptake as they realized quite early that if academics adopted the mediated deposit the proposal was unsustainable.

> And also once you started to work with the project and do the calculations you realize how unsustainable perhaps offering mediated deposit would be for an institution this size. *[CS5]*

Interestingly once the repository was in place the type of participation that they did get was mainly through self-deposit.

> Right at the very start there was some thought that some mediated deposit might be required to get people going but in fact as the project unfolded it was pretty much all self-deposited items. *[CS5]*

This repository manager concluded that the mathematics of mediated deposit does not necessarily work out for a large university. Additionally, promoting self deposit is considered much better as the IR then forms part of the research culture and is therefore more embedded within the way academics work.

The fourth repository [CS4] is loosely derived from an old repository that was part of a nationally funded project to promote repository development. However, when the funding for this project concluded, the repository was discontinued. Some of the material in the new repository is taken from there, although they have developed a very broad range of techniques for acquiring content including but not limited to university publishing house, subject repositories, self and mediated deposits and projects. They describe their collection process at the beginning as going for the low hanging fruit. This approach is used to gather as many items as possible and work towards achieving a critical mass of items.

> Mainly us going out [to talk to people] because it is new. To go out and tell people it exists. Nobody knows it is there at the moment really. So our assistant she is on the case and basically trying to spot low hanging fruit to get in it, anything we can get we are allowed to put it in. Also conference papers, grey literature we should put in. We now have a mandate for theses (…) We are also trying to get postgraduate students engaged so that those that are existing students will voluntarily submit. *[CS4]*

What they are particularly interested in is creating not only quality assured metadata but making it much richer than most of the other repositories. They are working on sophisticated metadata. This requires that repository staff spend a great deal of time creating metadata for the deposited objects. However, the repository manager believes that this is a key part of the repository and useful for dissemination and long term access.

A fifth repository [CS7] was created as a result of an internal project to look at repository development and was issued as a recommendation. Their initial collection strategy was focused on acquiring materials that were related to the schools and departments of the members of the project team or steering committee and based very much on personal connections. They then worked with key schools and increased their contacts. Currently they have mediated deposit and all items are checked by one person before final deposit in the repository. They are currently not actively looking for submission because they have a large backlog of material waiting to be checked before deposit.

> It is worth saying that the reasons that we have so much material that we are not going out actively seeking is to some extent because we built up a very large back log. We did quite a successful publicity campaign almost at the start of last year but then because the repository librarian post was vacant for some months, actually a large back log of material built up and there were other reasons as well why we were not getting things in and so to some extent we are still working our way through the results of that which is why we are in the fortunate position of having too much material. *[CS7]*

Another repository [CS2] stemmed from a bibliographic database used by the department, focusing mainly on journal articles. They were a very early repository, at the forefront of the OAI-PMH and Open Access developments, and they added full text to their entries. This was done mainly through self-archiving and they were one of the early departments to have a departmental mandate for submission. In the beginning there was not central quality assurance for metadata and deposits rates are very high.

The last case repository [CS1] was created from a file storage system. The items collected had no type of metadata but it was important that these digital materials were not lost. Funding was received to work on a system, specifically a repository, that could handle, organize and preserve these materials with the necessary metadata in order to store these digital materials in a more

useful and practical way. The initial material came from this file storage system and the following collections tended to be developed from personal interest or connections from the repository development team:

> Archaeology arrived when [name deleted] was here and that was actually because he knew someone that worked in archaeology who had some data, someone going with whom he went down to the pub. *[CS1]*

Indeed the criterion for accepting materials was related to its vulnerability rather than the importance of dissemination and/or access. One particular collection for example, was accepted because the researcher had an old computer, was a leading authority in her field and the data was in danger of being lost. Depositing it in the repository was a means of safe keeping the data. The rationale is that certain materials, such as journals and books, are in no danger of being lost as several groups are worrying about their preservation elsewhere. The burning issue is the digital items that are being created within the university by researchers or research groups and that if they are not deposited in the repository will in most likelihood be lost, either when the researcher moves on or changes computers or simply due to some technological mishap.

The types of materials within this particular repository therefore varied greatly. Organized by communities and sub-dived into collections there is no apparent method in the collecting policy. Many of the communities were developed from personal connections or interests. In addition they have created a number of communities that are empty and are acting simply as place holders. However, the remark from one team member to another during the interview: "is that project still running?" [CS1], is indicative of an unsystematic approach. The repository team is well aware that this has to be reorganized. However, they mentioned that one of the greatest difficulties they have run into is the politics involved with creating, deleting or reorganizing the hierarchy and distribution of communities and collections.

> It is a dark archive with video in it and it shouldn't be there. Frankly, it is too big but there is all kind of politics, a lot of this, some things have political repercussions, several times now. *[CS1]*

One role of publishers has always been to select and decide what is worthy of publication and what is not. One of their methods for sanctioning material is through peer review. In the case of this repository, the repository staff has inadvertently acquired a position of power, responsible for deciding what material may be deposited in the repository. This has been accompanied with a series of political issues.

Repository objectives

In the survey respondents were asked to rate the relevance of six statements about repository functions and main objectives. The statements and the rated relevance are show in Table 22. These drivers were chosen as they were key issues that were picked up in the literature review about the different functions that an IR can fulfill.

	Highly relevant	Relevant	Slightly relevant	Not relevant	I don't know
Enhance access	78	15	2	0	0
Encourage new forms of peer review	4	16	35	36	2
Encourage new modes of publication	27	32	24	13	0
Aid institutional information management	26	35	21	9	3
Promote data sharing	40	30	19	4	1
Preservation of digital resources	44	35	14	3	0

Table 22 - Percentage of respondents on repository function

The most relevant statement, according to repository objectives, was using them as a form of enhancing access to resources with 78% considering this highly relevant. Promoting data

sharing was also considered highly relevant by 40% and relevant by a further 30%. The preservation of digital resources was also quite high in terms of relevance with a combined total of 79% considering it highly relevant or relevant. There was less agreement on using the repository as a means of encouraging new modes of publication with 59% considering it highly relevant or relevant but 37% considering it slightly relevant or not relevant. The same holds true for aiding institutional information management because although 35% considered this relevant, 21% considered this slightly relevant and 26% highly relevant. Encouraging new forms of peer review was considered not relevant by 36% and slightly relevant by a further 35%.

The findings from the literature review indicated that the main drivers for IR development are for self-archiving, revolutionizing scholarly publishing and as digital infrastructure. Access and dissemination are key issues for all three types of IRs whilst preservation is relevant to the latter two. Finding that access and preservation are the two most relevant issues for repository managers ties in with these results. Additionally new modes of publication is particularly relevant to IRs that are hoping to modify the current scholarly publishing systems.

When analysing the interview data, although this was not a specific question, repository objectives or functionality emerged as a major theme as repository managers described their repositories. So although repository managers were not specifically asked what the functionality or objectives of the repository were; two repository managers mentioned preservation, three stated dissemination and access and one manager commented on the shifting objectives of their repository. Secondary objectives mentioned were curation and proof of concept from departmental to institutional.

Preservation

In the survey data the issue of preservation also came up in the open-ended questions. For repositories that are interested in handling a wide variety of electronic resources it is not only a

question of how these can be handled but also how can these be preserved in the long run. As put by one respondent:

> We accept any sort of electronic resource in our repository, which is great, no problem with that, but even though we only pledge to preserve certain formats... I wonder about having other, out-dated formats in there (we actually have one or more hypercard stacks). Is there a point to that? hmm. And even the formats we say we will "keep available" - how are we going to do that, really? I have not heard a real answer, so that's very scary... when people ask about that.

The preservation of non-formal resources was of particular concern. It seems to be that repository managers are aware that formal journal publishing is being preserved elsewhere and that the digital resources that are much more at risk are the ones that are outside the formal framework of digital publishing, such as datasets and grey literature. These are the types of materials that in the long run would benefit if placed in a repository for preservation.

> We are finding that more and more faculty are concerned about data sets - especially small to mid sized data sets - because they can see the problems with preservation. They don't tend to find the same argument about their peer-reviewed published material compelling. Data sets are, of course, complicated for libraries to deal with. Also departments see the logic of the repository service for their grey literature in particular.

In the case studies, two repository managers mentioned preservation as one of the main drivers of their repository. Interestingly though one case study [CS1] deliberately did not consider access and dissemination a driver, only preservation. According to the interview data it was not an objective because the university already offered a variety of web publishing services for this purpose. This was a particularly surprising find as almost all IRs, based on OAI-PMH have

been created specifically for dissemination. It could actually be said that this is almost a defining characteristics of an IR.

However, it should be added that for this particular interview three repository team members were present. After one member of the team mentioned that dissemination was not an objective another one added:

> That being said though, there is less management in those services because obviously people run them themselves and have less of a perspective from an information management focus. So we still want the scholarly content that, papers and what not, but our perspective is a lot broader than the sort of scholarly communication perspective that a lot of IRs have. *[CS1]*

It seems that even within repository teams there is not necessarily a consensus on the functionality of the repository, as in this case where one member of the team focusing mainly on archival and preservation issues whilst the other included access and dissemination as well. In addition, as this particular repository is organized by communities and sub-divided into collections, each one may also have different particular objectives. So for example, some communities can have double objectives.

> It is (the community) partly to make it publicly available and partly to have a digital preservation. *[CS1]*

This of course does not mean that the different drivers are incompatible, however in order to work toward building evaluative frameworks for IRs these different drivers must be taken into consideration.

Encourage new modes of publication

Some repositories are also looking, not just to store and to provide access to resources, but to function as publishing systems. Many repositories are already used for publishing journals. In some cases repositories are actually being used as a publishing space for different types of materials. One survey respondent mentioned:

> We view the repository software as a means of providing a university press to our clientele. We expect it to be a disruptive and subversive technology.

A case study repository manager indicated that by making information available they were contributing towards new methods of publishing.

> I think we are slowly evolving towards new methods of publishing alongside new ways of making information available. *[CS5]*

They were also, through the repository, changing the role of the Library. They no longer felt that they were only a place where one "just ordered books" [CS5]. The repository manager felt satisfied as she felt that they were now more actively engaging with academics in a proper intellectual debate about publishing and communication.

Enhance access and dissemination

Although access and dissemination are key drivers for most IRs there are different approaches to how to make this material available. In some cases IRs are concerned with assuring that they have a copy of the item within their repository.

> I have always followed if we don't have a copy it may as well not exist. Which is why I am against just linking to item

> in other repositories and saying we have got it to people. *[CS1]*

In other cases repository managers are interested in making sure that they have the actual digital objects for certain items types (mainly e-prints) but for other types they are fine with just having the metadata. When pressed for an example a case study repository manager replied:

> Patents, book, so there are various classes that we don't worry too much about. *[CS2]*

On the other hand, in other cases where the objective is to capture a representation of the university's research output and although they prefer to store the digital object they are willing to accept the metadata only as they do not want to exclude registering materials from the repository.

> A scope of records where we have metadata only but plus we have a link to an OA document elsewhere, so it might be BioMed Central for instance or it might be a report which is full available on another website but for whatever reason they don't feel able to put it in the repository. And indeed if you truly believe in OA should it matter? As long as it is available somewhere that is the goal. *[CS5]*

One case study repository manager argued that, for example, for the Arts and the Humanities it would be very difficult for them to obtain the copyright to store the full text of books and so they accept the metadata only record or a link to a section or part of the book.

Repositories that mentioned access and dissemination as important drivers tended to be the ones that focused on using the repository to showcase the university's research output. However, most of these repositories had had to make important changes to their policies as more demands were made on the repository, in particular as added on services. A strong case in point has been to move from only accepting full text items to settling for metadata only.

> We just reached a point where we had to make a pragmatic decision and recognize that a lot of material that was coming in we couldn't put it in full text (…). The feedback that we were getting from the university was that they were just as interested in [the repository] as a way of promoting research and providing links [to publishers websites] as it being pure OA. *[CS7]*

One important turning point for a few repositories was when the university decided to use the institutional repository as a register of all university research output, for example to submit RAE returns. The knock on effect has been that repositories that may originally have been aiming for full text have had to take metadata only records in order to ensure that the registry is complete. They have been:

> Taking away the emphasis from OA and more into [the repository] being a registry of university research. *[CS7]*

Another example has been offering the repository as a tool to automatically generate homepage publication lists. One repository [CS5] found for example, that after implementing this tool as a means to encourage academics to deposit their materials, they had to modify their policy of not accepting material that had been created at another institution. Academics wanted to be able to include all their publications and this meant that this adjustment was made. It could therefore be argued that the repository now represents the research output of the university and research conducted elsewhere. This leads to important issues about interoperability between repositories and multiple depositing.

Shifting priorities

One repository in particular mentioned problems with shifting priorities and drivers depending on organizational priority.

> It is a very hard question to answer. It seems like an easy one, why is your repository here? But you would be amazed, you can talk to all kinds of senior people and get different answers on that, you know there are lots of different perceptions on what [the repository] is there to do (…) and the perception has changed depending on what day you ask people. Or what week, what month, what their current priority is, what their current concerns are. *[CS7]*

Although to a lesser degree another repository [CS3] also mentioned the problem of defining and explaining repository functionality in order to train other members of staff, such as a librarian, to help with the advocacy. The example he gave was overhearing a librarian explaining to a member of staff the role of the repository and noting a number of misconceptions and errors. Repository advocacy is complicated with different actor's understanding of the role of the repository according to different roles within the scholarly communication chain and their agendas.

It seems to be that repository functionality is for all repositories still in flux, if not on paper, definitely so in practice. This can help to explain the lack of agreement and the inconsistencies in the genres and types of materials that are accepted within the repository as well as apparent contradictions.

Usage and visibility of electronic resources

Perceptions on usage

In the survey, repository administrators were asked if they considered that the materials within their repositories were used frequently. This question was a runner up to more detailed information about how they knew this. Over half (53%) of the respondents considered that their

materials within their repository were used frequently, with a further 25% stating occasionally, while close to 4% said rarely or never, as shown in Table 23.

Resources in my repository are frequently used	Frequency	Percent
Never	1	0.7
Rarely	5	3.4
Occasionally	36	24.8
Frequently	77	53.1
Other	5	3.4
I dont know	21	14.5
Total	145	100

Table 23 - Use of electronic resources within a repository

It is interesting to note how over half of the repository managers were quite confident in asserting that resources were well used. However, almost 15% replied that they did not know. From the open-ended survey responses it would seem that some repositories are aware of this issue but are only just setting up the facilities. Under staffing is an issue for repositories and most have concentrated primarily on gathering content.

> We are just now developing facilities for the capture and reporting of usage statistics and are evaluating a number of open source tools for this.

Usage monitoring

Respondents were then asked if the use of electronic resources within the repository was monitored and if so, by whom (see Table 24). Almost 90% of respondents indicated that resources were monitored, with 12% indicating that they were not monitored or were not aware of this. Over 35% claimed that they were involved in monitoring usage data, whilst a further 24% reported library participation. A special committee or computing services were also involved (14% and 12% respectively).

Monitoring groups	# of responses	As a percentage of total responses
Computing services	28	12.6
Library	54	24.4
Myself	81	36.6
Special committee	31	14.0
Not monitored	24	10.8
I don't know	3	1.3
TOTAL	221	100

Table 24 - Groups in charge of monitoring repository usage

Methods for monitoring

Respondents were asked how this usage was monitored and given several options: download counts, Google analytics, link analysis, log analysis and user surveys. A free text option was also available.

Of the respondents who monitored usage data 41.9% used transaction server log analysis. A further 35.8% relied on download counts, followed by Google analytics with almost 14%. User survey and link analysis were the least popular methods (see Table 25).

Method	# of responses	Percentage of total responses
Download counts	82	35.8
Google analytics	32	13.9
Link analysis	8	3.4
Log analysis	96	41.9
User surveys	11	4.8
TOTAL	229	100

Table 25 - Methods for usage monitoring

Some remarks showed a very detailed knowledge of what particular resources were used frequently and even by whom and for what.

> Everything in the repository gets used a bit, but collections such as the musical scores get used A LOT. Our 3 highest use sets of items are the musical scores, computer science tech reports, and a collection of student papers from a course - (about researching gravestones). The latter are used by people doing genealogy research.

Others indicated that they were aware that usage statistics were important but that this facility was still in development.

Use of usage

Data from the interviews indicated that six out of seven case studies reported keeping usage statistics, specifically downloads, but all noted that they were extremely basic. Repository managers reported however that these statistics were valuable in particular for proving the repository's usefulness to funders, steering committees, advocacy and promotion. The following selection of quotes illustrates this:

For the funders:

> Well people ask: why should we give you money? Then you need graphs. That's what I use them for. *[CS1]*

For steering group:

> For the first time in the steering group I have produced sort of pretty pictures with stats because they have asked for them. *[CS4]*

For advocacy:

> They are nice useful tools that you can use for advocacy, useful things to prove to panels that you want to get funding to go on. *[CS3]*

For promotion:

> But to be honest they have been used as a bit of a promotional tool, that is the main thing (…) mainly been a way of reporting back to senior management and our steering group as an indicator of project success. *[CS1]*

Repository managers were essentially interested in usage statistics:

> Usage, I have come to realize, is a very interesting piece of evidence about display of scholarly impact. *[CS2]*

Issues with usage statistics

Repository managers were aware however, of the difficulties of interpreting download data. One repository manager argued that if traffic volume is equivalent to quality then you can "play the game by just using a rich vocabulary and attract as many queries as you can". [CS2] Although download statistics can be a useful indicator, in particular for justifying the repository's usefulness, mangers are aware that this still does not provide any answers about what people are doing with the resources that they are apparently downloading:

> I am speculating based on reasonable evidence (…) I don't think anyone has very much concrete evidence why or how people are using things. *[CS5]*

One repository manager discussed in length the conceptual differences between counting abstract downloads and full text downloads. He was referring specifically to e-print downloads, which are quite regular in structure and for which an abstract and full text are usually available. For non-textual materials this of course is further complicated as they will probably have a more heterogeneous nature. When counting download what is the difference between downloading an image, a full-length video or a podcast? How can the implications of their use be quantified in this way?

An additional complication with usage statistics as downloads are items that are actually made up of several discrete files and together make up an individual item. An example of this would be audio files with an interview transcription, included under the same metadata file. This issue is fairly similar to the one encountered when attempting to detect links to an item in the link analysis, which is discussed in the link analysis section. If downloads for articles which are fairly regular in their structure are problematic, measuring downloads for complex objects will of course prove more challenging.

A similar issue with identifying units of analysis also surfaced in relation to metadata only records and full text items. For repositories that do not necessarily require the digital object to be deposited but will accept a link to the object contained in another database or repository (freely or not) this would signify another important conceptual difference.

Other usage indicators

Repository managers were also asked if they had any other indications of repository resource use alongside download statistics. One repository manager described that on one occasion when the IR server had been down they received an email from Google Scholar requesting information about the situation. This appears to indicate the repository is an important source, probably due to the sheer volume of resources available, for the search engine. It may also indicate that there is a demand for 'scholarly resources' that are identified within the enclosure of a repository. However, in this particular case the primary function of the repository was preservation. So although they were pleased that the resources were useful their primary interest was not external use.

A case study repository manager mentioned that they received emails from individuals thanking them or indicating that certain resources had been useful for their work. Again, when a server was down, emails had also been received requesting further information as these resources were used frequently. However, it seems that little importance was assigned to this type of indicator

of usage, mainly because it was more qualitative and not quantifiable: "Well we enjoy that but it is not really quantifiable" *[CS6]*. Repository managers tended to mention these types of indicators more as anecdotal evidence than concrete usage evidence. Interestingly enough the answer to what resources are actually used for could probably be found in these types of indicators rather than in large, quantitative download statistics.

Link analysis

The link analysis was designed specifically to look at visibility of the resources within the repositories. As results two typologies were developed: one of the target items within the repositories and the second of the source pages that linked to the items within the repositories and these are presented. We then look at the types of target pages linked to and the source pages linked from according to the figures provided by the link analysis.

Target page typology

All seven case studies had the target types described within the metadata fields of the item and these were used as the basis for target typology. Although all repositories included an item type field in the metadata, classification schemes varied slightly and there appears to be no standardised inter repository system yet. In some cases, the labels were quite similar and conceptually almost interchangeable. For example, book chapter or book section. However, in other cases, although the label described a similar item type (for example, an article) the different naming schemes were conceptually different (for example, preprint or postprint). This is particularly relevant if we consider their differences in the light of their possible role in the future of scholarly communication and publishing. So although, preprint and postprint are both articles, the distinction within the repository suggests a conceptual differentiation. Similarly some repositories added as a sub-field or within parenthesis the item's publication status, again conveying a notional difference for a published or not published research report, for example. Other repositories chose to use a more general top-level label (for example, monograph) and use

a sub-field to refine (for example, research report or working paper). Other repositories had a wider classification system and would have a working paper as a top-level field.

Using this particular system the initial list of target types included over forty categories. This was considered too long, especially regarding the fact that some variations were quite slight. It was considered important to group similar labels beneath one umbrella category. In some cases it was necessary to ignore the detailed information (such as a chapter in a book) for the more general category of book. This was done in order to homogenize all the items listed. Table 26 shows a list of all the initial target labels with a short description of their differences, and the final grouped target category assigned.

Target types	Category differentiations	Target categories
Article Article: in press Article: postprint Article: preprint Article published Pre prints Journal article	Status of publication	Article
Book: published Book Book (monograph): Section of book or chapter of book Book chapter Book section Book section: In press Book section: Published	Nature of division (whole book, chapter or section) Status of publication	Book/Book chapter
Monograph (discussion paper): Published Monograph Technical Report Technical Report Monograph (research report): Published Monograph (research report): Unpublished Monograph (working paper): Published	Type of monograph (research report, discussion paper, working paper, etc.) Status of publication	Research report Technical report Working paper Documentation Monograph unspecified

Working paper Monograph Type Working Paper Monograph Type Documentation Monograph Type Unspecified		
Conference item (Paper): Published Conference item (Paper): Unpublished Conference item (Poster): Unpublished Conference item (Presentation): Unpublished Conference or Workshop Item Conference or Workshop Item (Lecture) Conference or Workshop Item (Paper) Conference or Workshop Item (Poster) Proceedings paper: In press Proceedings paper: Published Presentation	General types: Conference item, Workshop, Proceedings Types of items: Lecture, Paper, Poster, Presentation Status of publication	Conference
Thesis	Thesis	Thesis
Recording, oral	No mention of what (lecture, conference?)	Audio
Video	No mention of what	Video
Other	Files in Chemistry Mark up Language (.cml)	Other
ADDITIONAL		
Code Email link	XML pages for OAI-PMH requests (identify, get records)	Code
Collection Community Page Search Page	Groups or lists of items	Browse
About	About repository page, help page, FAQs and similar	About

Table 26 - Target types, characteristics and final categories

A number of the links found were not to a particular item but to other sections of the repository. Some repositories offer items grouped by community or by collection (for example, the Social Anthropology collection or Conrad Martens' Sketch Books). Other types of links were to a search query, for example an academic's last name, giving a list of all items by that author within the repository. These three types were collected under the general label of 'browse', as they were linking to a particular group of items. In these cases the link was noted but no information about the actual items listed was registered.

Another group of links found were to XML pages within the repository that answered queries, usually OAI-PMH requests. For example, identify, list records, list sets. These types of links came predominantly from repository directories. There was also one link found to an email address. These types of links were grouped under a 'code' category. As with the previous example of browse, no information for an item was registered. The category 'about' covered links to pages within the repository that described or gave further information about the repository itself, including policies, description, background information and so forth.

The final list of content types is very extensive and it is clear that developing digital content typology is a complicated issue. As mentioned previously, in the online survey repository managers were asked to mention the types of documents that were contained within the repository by selecting from a list taken from OpenDOAR. Table 27 shows a comparison of the initial OpenDOAR, new OpenDOAR and the link analysis target page typology. There is little variation among the category types and non-formal resources are underrepresented.

New OpenDOAR	Old OpenDOAR	Target page
Articles	Preprints	Article
	Postprints	
Books	Books and book chapters	Book/Book chapter
Conferences	Conference proceedings	Conference
Datasets	Datasets and databases	
Learning objects	Learning objects	
Multimedia	Audio-visual materials and multimedia	Audio
		Video
Patents	Patents	
References	References/bibliographies	
Software	Software	
Special		
Theses	Theses and dissertations	Thesis
Unpublished	Reports	Research report
	Working papers	Technical report
		Working paper
		Documentation
		Monograph (unspecified)
	Administrative documents	
	Images, maps, diagrams	
		Other
		Code
		Browse
		About

Table 27 - Comparison of old and new OpenDOAR and target page typologies

If we take into consideration that developing digital document classification systems is a difficult issue in itself and furthermore that an institutional repository one has not been attempted before, then general target typology can serve as useful start from which to continue working. The resulting classification list, though general, is useful enough to permit analysis on the item types collected within repositories.

Source page typology

Although not the main focus of the research, a source page typology was developed in order to provide a framework for looking at the target resources. What types of pages are linking to items within institutional repositories? As mentioned previously a few classification systems exist and these were taken into consideration for the development of one for this research. When

using these lists to attempt to aid classification it became clear that the older classification systems (from 1998) were difficult to apply. For example, the use of homepages that were popular in the 90's, has evolved and divided into more precise, identifiable genres. The homepage as such is still present in the form of, for example a researcher or academic homepage, but some of the homepage functionalities have been taken over by other types of genres such as social networking sites (ie Facebook or MySpace) that are a new genre in themselves. In addition, new genres have evolved, for example wikis or blogs, which are not considered in these older category listings. Others such as list serv mailing lists and discussion forums are still very much alive and in the same spirit as several years ago.

Digital classification systems are a moving target and the list presented here will most likely need adjustments in the near future. However, attempting to classify digital documents is still an important exercise in particular because they help trace the development of emerging digital genres over time. In addition as new genres are accepted and incorporated these will become a recognizable genre in digital typologies, as is becoming the case with for example blogs. Additionally if we are to understand the impact of new digital resources on scholarly communication and publishing we must be able to, even if only generally, classify and name the different types of documents that we are discussing in order to avoid treating all digital resources as the same.

Defining the page unit

The first step is to define the web page unit to classify. As mentioned previously this does not have one simple solution and the problem has been grappled with by other studies (Crowston and Williams 2000; Thelwall 2003a). One example of the difficulties with defining the unit to classification would be to take an HTML page that is a reference list at the end of a paper of a conference of a particular organization's webpage. Depending on the level of granularity of the classification system and the extension of the web page unit the source page could be a reference/bibliographic list, an article, a conference page or an organization page.

For this study, as with the target typology, the source page classification system was designed to be general rather than specific. Therefore a bibliographic reference list of a conference paper would be classified as conference. However, a classification system that was too general would not be useful either, so over generalization was avoided. So for example, a researcher's homepage was not classified as a university page (and this category doesn't actually exist) but rather as an academic homepage. The target page typology developed is presented as Figure 24 followed by a brief description of each source page characteristics.

Academic homepage	Government/National body	References
Blog	Indexed/Search engine	Repository
Company	Journal/Magazine	Repository directory
Conference	Mailing list	Research project
Course	News service	Topic
Departmental homepage	Organization	Unknown (foreign language)
Discussion forum	Other	Wiki
Dot com linkpage		

Figure 22- Source typology

Academic homepage

A web page describing an academic or researcher's profile, usually including research interests, publications, CV and contact information. These pages were regularly hosted under a university domain but a few self-hosted academic homepages were found. The links to repository resources most frequently came from the academic's publications lists although a few exceptions were found. For example, one researcher used a link to a paper within the repository in order to expand a particular theory that he described in his research interests.

On several occasions links to the same target item were found, in particular with doctoral thesis. These tended to come from an academic who created a homepage at two different institutions, either because he or she had moved or was a visiting professor. This exemplifies an important topic currently discussed about the relationship between institutional repositories and subject

based repositories and how to deal with academics who will probably move to several institutions during their careers. It also poses an interesting question for link analysis in terms of link counting.

Very few student homepages were found and because of this a separate category was not created but rather they were grouped under the category of academic homepage.

Blog

Blogs have become a recognized and easily identifiable genre with known characteristics. Blogs allow usually one author to submit what are known as posts, and these are presented in chronological order. A large majority of the pages identified as blogs were theme based and discussed issues surrounding a particular subject. Although the source page category topic was also created, it was thought that blogs, due to their very particular characteristics, should be placed in a separate category. In particular because blogs are a very distinctive digital only genre and categorizing them separately could shed some light on new forms of scholarly communication and publishing.

Company

This category was assigned to web sites that belonged to a privately owned for profit enterprise. The websites presented information about the company, staff, contact details, among other information. An example of the types of links found are a company who linked to several papers that had used their products in the research.

Conference

A web page dedicated to a conference, usually including conference details (venue, dates), practical information (hotel, maps), presentation details (program, speakers) and papers (abstracts, full papers, presentations, posters). On occasions conference web pages were hosted under a university domain but not necessarily. The links to repository items tended to be either

from conference references in a paper, or from a conference page that had links to the presentations/posters/abstracts. Workshops were included within the conference category.

Course

A page giving details about a particular course (summer course, university modules, diploma, seminar) usually hosted under a university domain. The links to the repository items were usually found in reading lists, references or bibliographies or as part of final assignments.

Departmental/Centre homepage

A web page belonging to a department or centre within a university, usually offering information about the department including teaching, research, staff, application procedures, information for current students and publication lists. Links to repository items were from departmental student theses lists, the homepage itself or from the departmental publications list.

Discussion forum

A fairly established digital genre, discussion forums are usually topical and have several participants. The links established to repository items were usually found when arguing or exemplifying a point. One example, is a forum discussion on the nature of evolution linking to an academic article in the repository as evidence for a particular point.

Dot com linkpage

A web site that has a long list of links, apparently generated randomly whose only function seems to be to generate traffic, spam search engines or other type of non-content activity. When these pages were analysed the links to the repository case study were no longer available. Considering that these pages tend to automatically create links and update frequently the time lag between collecting the links and analysing the web pages was too long. Due to the way that they are created these links are not particularly useful or meaningful.

Government/National body

A web page belonging to a specific government or national body (for example, the Institute of Marriage and Family, Canada). Although this could be considered a rather general category, there were very few cases this type of source pages (four) and therefore no further sub classifications were developed. The links to the repository were from the resources/references or publications section of the web sites.

Indexed/Search engine directory

This category refers to pages with list of links that are created either automatically or by humans but are done so in a more targeted and relevant manner than the dot com linkpage. These link directories are a more meaningful resource as most of the links are placed under relevant subject categories which have direct reference to the items that they are linking to. For example, a link to a video on the Industrial Revolution was found under the Industrial Revolution category.

Journal/Magazine/Ebook

Surprisingly very few of these types of source pages were found and it was therefore decided that they should be grouped together. The links to repository resources were usually found in the reference or bibliography section although there was one case of the link being found within the actual text.

Mailing list

These are messages from mailing lists that have been archived online. The links to repositories were usually to illustrate a point, add further information or as dissemination by announcing the availability of a particular resource.

News Services

Newsletters, newspapers and RSS feeds were all grouped in this category. In general the category refers to web sites that are primarily dedicated to offering news items, rather than a web page that includes as part of the its general content, news items. So for example, if an

organization page offered news on their home page this was classified under organization rather than news. For the news service category it had to be the primary function of the web page.

Organization

Web pages for organizations that are apparently not for profit and are usually working on a particular academic subject area or topic. Examples are the Council for British Archaeology or the Institute for African Alternative. The links to repositories tended to be from the references or publications section of these web sites.

References

This category includes bibliographies or web bibliographies that did not belong to a conference, department, academic or another category page or were provided more as an independent resource (for example a Library page with links to Open Access information). This category is also for bookmarks and citation indexes and other tools for managing and sharing online resources, such as CiteULike, Citebase or SnipIt!

Research project homepage

Web pages created by a particular research group or project in order to present information about the work undertaken either as it progresses or as a final dissemination medium. These pages tend to be created under a university domain or hosted under a specially created domain name. Most research projects were either within a university, inter university or with private consultancies and/or government bodies and universities. The links to the repository tended to be within the publications section of the web site.

Wikis

Wikis, like blogs, have an easily identifiable structure and underlying technology that makes them easy to recognize. However, wikis are less of a genre than blogs and can be used for creating different types of pages. For example, wikis can be used for creating conference website. In this particular case, it would be classified as a conference as the focus on the

classification schemes was more on content or purpose than underlying technology. An important defining characteristic for a wiki is that usually the intention is that it will be edited by more than one person. Wikis are particular popular with groups of people working together on projects and documentation and are frequently updated.

Other

The category other includes three different types of web pages that could not merit an independent category but are interesting and particular enough not to assimilate them under another more general category. The first are code pages, rendered in HTML of XML, usually with links to XML files within a repository. The second is a file sharing site although this had no link. Finally a cruise information data set. This one could have been considered a centre (Marine Seismic Data Center) but because it was providing raw data, which none of the other pages had, it was considered best to keep it separate as a possible case of novel types of web publications.

Repository

The category repository is fairly easy to identify as they are fairly recognizable offering metadata information and deposited items, search and browse facilities, similar software (Dspace, Fedora or Eprints), support OAI-PMH and tend to auto-identify themselves as such. There were quite a few cases of repositories as source pages. Usually the link to the case study repository was as alternative location for an item that they had registered.

Repository directory

A website offering a listing of repositories usually at a worldwide level. Some can be searched and browsed by name, country, content type and so forth. Usually the links to case repositories were to OAI-PMH identify requests.

Server page

This category gathers source pages that were either blank, file not found, page not found or a welcome to Apache page. These sources pages obviously have no target links. The page was probably moved, or the server set up again and the link no longer available.

Topic

A website dedicated very specifically to offering or developing information on a particular topic. For example, a web site dedicated to the history of the E-journal or Transport and the Environment.

Unknown (foreign language)

This category groups the source pages in a foreign language that were not easily identifiable and language barriers did not allow for a more precise categorization (a clear example is pages in Japanese or Arabic). It was considered better not to attempt to classify them. Not all pages in a foreign language were classified under this category as some of these source pages were identified by looking at their structure or URL. For example, blogs are fairly regular in structure or departmental homepages could be identified through URL and structure and this were assigned the corresponding category.

Target pages linked to

After classification, the target data for all seven case repositories was analysed. Table 28 shows the types of target pages and the number of links found for all case repositories.

Target type	Num of links	As % of total
Article	122	29.19%
Book	75	17.94%
Conference	57	13.64%
Code	24	5.74%
Technical report	21	5.02%
Video	21	5.02%
Browse	19	4.55%
Research report	17	4.07%

Thesis	13	3.11%
Not found	11	2.63%
Working paper	11	2.63%
Other	7	1.67%
Image	4	0.96%
Mono unspecified	4	0.96%
Discussion report	3	0.72%
Search	3	0.72%
About	2	0.48%
Audio	2	0.48%
Documentation	2	0.48%

Table 28 - Target pages and links

Articles were the item type most frequently linked to, followed by book/book chapter and conference. All these item types belong to the more established formal digital publishing realm and there appears to be a continued preference to use these types of resources. It could be argued that repositories contain more of these types of materials, thereby increasing the probability of linkage. From the survey results articles were the most frequent type of items within a repository confirming this assumption but conference proceedings and book chapters did not figure among the top three most frequent resource types. Reports, working papers and theses were more frequent. One of the major arguments for including theses within repositories has been to increase visibility of graduate research but this is not shown in the linking. Interestingly although book and conference proceedings are not as frequent they do receive more links. This seems to supports the finding that as with citation, in linking, there is still a preference for items that are more 'published'.

Other categories with paper based equivalents, such as research report, technical report, monograph and discussion report received a lesser percentage of links. However, the numbers vary as per case study, as show in Figure 23. One possible explanation is that these four categories have blurred boundaries and repository may have different ways of naming similar item types. These types of items in the print world belong to the grey literature realm and

classification schemes are less precise than for example with books or journal articles. If we merge all the categories together they account for almost 14% of the links, just below article and book.

Very few links to non-textual material such as video, audio or image were found. Although repository managers in the survey listed a wide variety of types of materials in the repository, this was not reflected in the links. It may be that users are not familiar enough with these genres to link to them. Another explanation is that these types of items exist in small numbers, as shown in the frequency results, and the link sample was not large enough to evidence the presence of these types of materials. CS1 for example, whose collection policy is much more orientated towards non-textual material was actually the case study with most links to image, video and audio. It therefore seems possible that there are no links to these types of materials in other repositories simply because there are so few or none at all. This supports the finding from the interviews that repository managers are stating that they would accept a wide array of materials but in practice these are not being deposited.

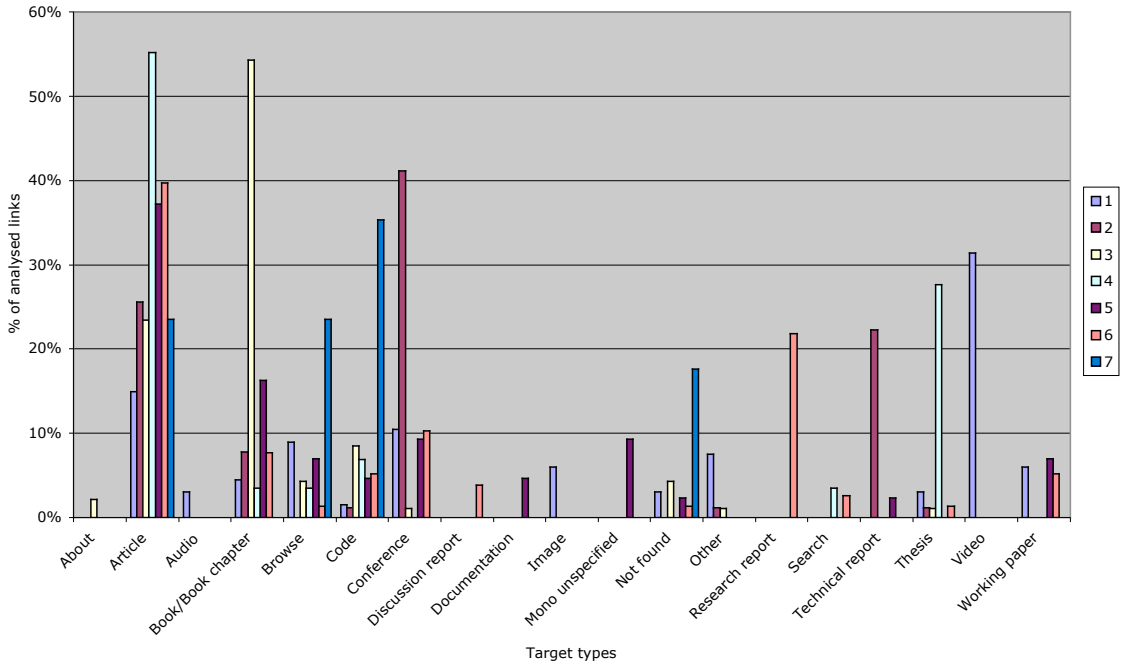


Figure 23- Target type by case study

Book/Book chapters were also a frequently linked to item from all case study repositories. One repository, CS3, received a disproportionate number of links to this type of item, over 45 links (48% of total analysed links) to book/book chapters. On closer inspection all 45 links are to the same target page and all 45 source pages contain exactly the same text including the link. It appears to be that one particular page (possibly from Open Directory project) has been copied by lots of different web sites (mainly dot com linkpages and index search engines), skewing the results. It shows how one popular text will build up and become even more popular. The repository confirmed that this was also the second most downloaded article and that additionally the full text was downloaded more often than the abstract that is the reverse of the phenomenon that they usually observe (CS3, personal email 08/07/08). Interestingly enough the links found were to the full text and not the metadata (abstract) page. Apparently high number of links can translate to high usage. A similar situation occurs with a CS1 target page. A Wikipedia entry for the Industrial Revolution with a link to a documentary video in the repository has been copied by numerous other web sites. 22% of CS1 links are to this particular resource. Both Wikipedia and Open directory project are licensed to be reproduced by others and multiple copies will create multiple links.

Conference was a popular category, particularly with the case repository CS2. This departmental repository is dedicated to the computer science and electronics fields that are known for relying on conference proceedings for scholarly communication. This suggests traditional publishing patterns are copied within repositories and academics continue to rely on traditional sources.

With a relatively lower percentage of links but both browse and code received similar amounts of links across all case studies. Other, in the CS1 repository, refers to a large collection of .cml files (Chemistry Markup Language) stored within the repository. This is one of the few evidences of linkage to new types of materials.

Source pages linked from

All source pages were analysed and classified according to the source typology developed. Table 29 shows the breakdown of source pages from all case repositories.

Source type	Num of links	As % of total
Indexed/Search engine	54	12.92%
Dot com linkpage	49	11.72%
Blog	47	11.24%
Academic homepage	40	9.57%
Discussion forum	28	6.70%
News service	23	5.50%
Research project	23	5.50%
Unknown (foreign language)	22	5.26%
Course	14	3.35%
Organization	13	3.11%
References	13	3.11%
Topic	13	3.11%
Repository	12	2.87%
Wiki	12	2.87%
Departmental homepage	11	2.63%
Conference	10	2.39%
Repository directory	10	2.39%
Journal/Magazine	9	2.15%
Company	8	1.91%
Government/National body	4	0.96%
Other	2	0.48%
Mailing list	1	0.24%

Table 29 - Source pages linked from

The indexed/search engine and dot com link page categories accounted for almost 23% of links to all repositories. However, dot com link pages are not particularly useful for link analysis as the links are generated randomly and there is little or no motivational reason for linking to a particular resource. Indexed/search engine links can be more indicative of perceived resource

usefulness, particularly very specialized indexes which are often important and frequently used tools for discovering online materials.

Blogs are the third most popular source for links to repositories. Links within blogs to repository resources tend to be focused and targeted, usually mentioned in the actual post text, either indicating the existence of a particular resource or using it as a reference for further information. It is worth noting that blogs, a specifically digital genre, are a significant source page for links to items within institutional repositories. This could indicate an important new publishing and communication tool both within academia and the wider public.

Academic homepages are another popular source for links to repository resources. It is actually surprising that these types of links are not more common considering that one of the functions of institutional repositories is to showcase research material. This could be related to the fact that university academics have not adopted depositing their research materials in institutional repositories as expected. How aware are university academics of their repository? Results from the interviews indicated that managers felt that their repositories were still not particularly well known within their institution. It has been noted that academics are reluctant to deposit and this trait may be evidenced by the low amount of links to repository items from academic homepages. An interesting exception is CS4, which actually has high percentage of links from its departmental homepages (14%) and academic homepages (52%). However, the CS4 repository is fairly new and only a total of 29 links were analysed so this data must be viewed with caution.

Discussion forums were surprisingly an important source for links to repository sources. In particular the links tended to be in the context of an argument, acting as a reference to support or expand on a given point. The discussion forums ranged from academic to more general audiences. This is an interesting indication of repository material being used outside the formal academic environment. An objective of institutional repositories is the wide dissemination and

use of academic material and this could suggest evidence of the general public using repository resources.

Research projects were consistently a popular source for repository linking from all seven case studies. Institutional repositories appear to be a popular place for research project to store their working documents and reports. In the print world these types of materials would generally remain unpublished and therefore difficult to find despite their usefulness for others. It has been argued that repositories are an ideal place for storing and disseminating this type of grey literature. Link analysis results supports this idea.

Wikis are not a particular common source for links to repositories but they are a consistent source across all seven case repositories. There is a wide variety of target pages for wikis however, making it difficult to draw any conclusions from the possible role of wikis in relation to resources within repositories.

Only two repositories registered news services as source pages. In the case of CS6, almost half of the links were to the same resource, which was a report that had only recently be made available. It is important to note the effectiveness of announcing the deposit or availability of an item within a repository in the appropriate channels for them to become linked to. When announced and disseminated a resource become more 'published' than if it just simply deposited. As described in the literature review one of the publishers' roles is to make material available to an interested audience. Although OAI-PMH is a useful dissemination technology using additional direct channels is even more efficient.

The news link in the case of CS1 is to a complex object. It is important to note, that unlike other repositories, many of CS1's resources were made up of several files (for example a text, an image, a zip file). In some cases links would be directly to one of the files but in most cases the link was to the metadata page that contained an array of different files. For analysis purposes

the first file listed would be used for classifying the target source. This is not an ideal solution. Complex objects should become more popular as repository systems become more sophisticated and the problem with defining the web page unit (in this case item in repository unit) will need to be addressed.

The category of unknown (foreign language) was quite high for CS1. Although the actual source pages cannot be defined further, it shows the use of the repository by other country's websites, possibly indicating a wider, more global impact of the research materials as made available through the repository.

Government and national body websites were relatively low on the list, with a total of four links in only two repositories. It has been argued that university research materials should be made available freely available, particularly when research has been funded by government or national bodies. It is interesting to note that currently there is no linkage evidence of these institutions taking advantage of the material available. As mandates for depositing government funded research materials become more popular this could change.

Company had few but evenly distributed number of links across the repository case studies. In most cases links were examples of research were their products had been used. A possible indicator of resources within institutional repositories having a greater online impact would be for these figures to increase.

It is particularly interesting to note that traditional print publications that have a digital equivalent, e-books, e-journals and e-magazines had very few links to repository resources. In the case of e-books and e-journals it could be said that although the medium has changed the style of writing has not altered, leading to traditional citation forms. It would have been interesting to find journal articles citing preprints within a repository for example. In the case of e-magazines, it would be interesting to see more popular dissemination using repository

material, following once again this idea of using repositories to make research materials available to a broader audience.

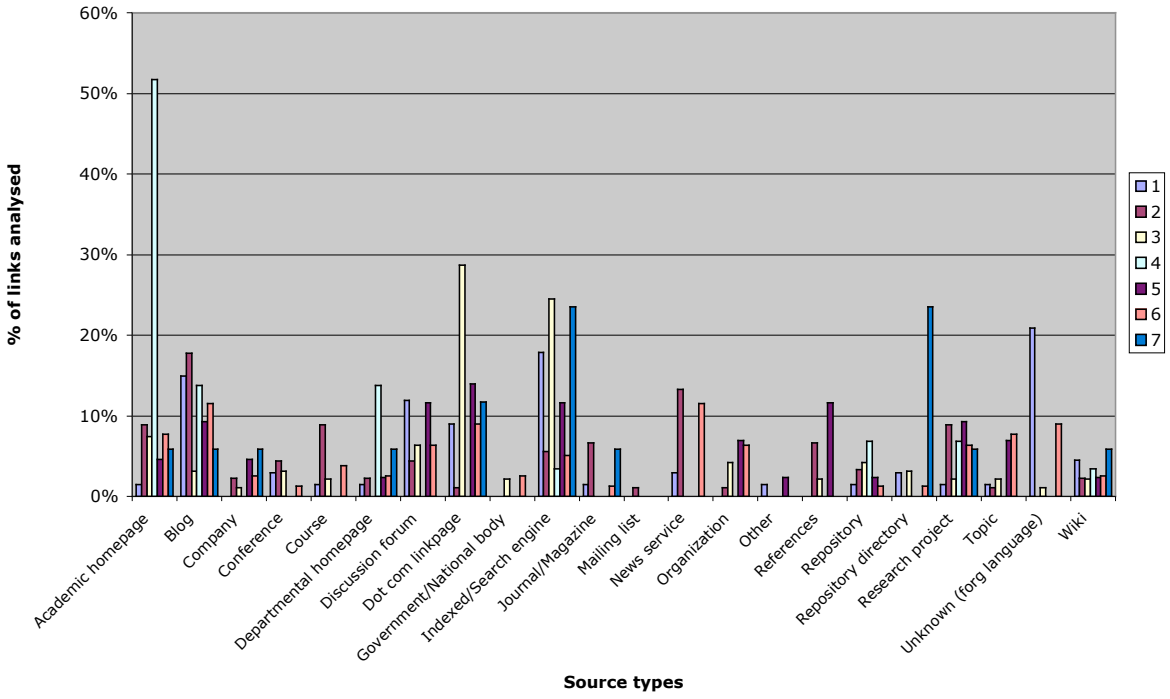


Figure 24 - Source types by case study

Overview and general remarks

In order to gauge opinions about several general aspects about repositories and scholarly content types, respondents were asked to rate their level of agreement or disagreement with five statements. The results are presented in Table 26.

	Strongly agree		Tend to agree		Neither		Tend to disagree		Strongly disagree		No answer	
	F	%	F	%	F	%	F	%	F	%	F	%
Books and journals are the only way to communicate research	1	1	11	7	8	5	37	23	97	**60**	9	5
A rep should contain a variety of electronic resources	82	**50**	49	30	12	7	8	5	4	2	8	5
Unmonitored repositories fill up with junk	8	5	43	**26**	23	14	23	14	32	20	8	5
Rep should only contain peer reviewed resources	10	6	10	6	18	11	44	27	64	**39**	8	5
New types of electronic resources will change scholarly publishing	61	37	73	**45**	8	5	5	3	7	4	9	5

F= frequency and %= percentage.

Table 30 – General statements on repositories

Open approach to acceptance of materials

Over 80% of respondents did not agree with the statement that books and journals are the only way to communicate research which backs up the idea of scholarly communication working with a large variety of scholarly materials and outputs, in particular grey literature and other non-formal resources. In addition a similar amount, 80%, agreed that repositories should contain a wide variety of electronic resources with a 10% in disagreement.

The open-ended question results also showed a general consensus towards allowing a variety of different electronic resources to be deposited in repositories. Some respondents indicated the need for a very open approach to this issue, in order to encourage new forms of communication, instead of limiting them.

> I think that we should not limit our options with regard to the types of publications and materials that a repository can hold. In many cases and in particular with learning objects, if we narrow our criteria we are reflecting a particular teaching formula which has not yet evolved toward understanding new teacher-student paradigms.

Some respondents felt, that although an open approach was the most desirable, some form of criteria for defining types of electronic resources should be in place. The selection of types of resources that can be deposited within a repository should reflect the objectives of the repository and be in line with some kind of a collection policy.

> This question really depends on the stated purpose of an archive. I think a clear mission for an archive should help drive the policies about who should submit and what should be included. But in general terms, I lean toward an inclusive philosophy, as it is both more democratic and likely to lead to a larger and better rounded archive.

> It depends on the purpose and aims of the repository as to what types of items will be permitted and who will be permitted to deposit. Clear policies are required to define exactly the scope of the repository.

Quality assurance and peer review

There is little agreement on issues of quality control of electronic resources and ways of ensuring this. 31% agreed with the statement unmonitored depositing would lead to a repository filling up with junk, with a similar percentage 34% disagreeing with the statement. However, almost 40% strongly disagreed with repositories only containing peer reviewed resources, and a further 27% tending to disagree, while only 12% agreed with the statement.

Some of the repository managers that supported the idea of repositories managing a wide range of materials added that this should be done within certain standards. There is a fine balance between remaining open to new forms but also assuring quality control.

> There needs to be quality control, largely driven by collection management, in the same way as libraries determine what goes in the library. But there also needs to be flexibility, at least initially, to aid familiarity with the system and help understanding of its role.

How these standards should be defined and evaluated was more difficult to define and there were different approaches among the respondents. Some mentioned using peer review and non-peer review as a way of indicating the quality of the materials within the repository.

> It should be made clear what the status of the material is: PEER REVIEWED or NOT PEER REVIEWED and WHO has DEPOSITED it.

Although there were strong advocates of using peer-review this was not a popular proposition with all respondents as illustrated by the 40% who disagreed with the remarks. Only catering for peer-reviewed items limits the uses of the repository.

> Exclusivity towards peer review removes the freedoms that a repository provides. Academics work in different ways and should be allowed to place works in progress in an online context.

Respondents seemed to want a very open approach to collection but at the same time were not sure exactly how to determine what can be useful. This is an issue that is still very much under discussion.

Metadata

One respondent made an interesting remark about the problem not being the variety of materials collected within a repository but exposing all the content and metadata in the same way.

> While I noted above that I tended to agree that depositing ANY form of content would lead to a 'junk-filled' repository I think that it's worth noting that one person's junk is another's gold. (…) I think that the idea of a one size-fits-all EXPOSURE of content is a questionable approach. Thinking that there is a single value proposition for IRs is flawed in my opinion.

In this sense metadata about the resources within the repository are important. So although there can be a wide scope for different types of materials it is important that there is contextual information which allows the user to identify and use the resource.

> The contextual information about the resource is important. It should be clear to the user what version of an item they are looking at and if it is peer reviewed or not. The aim of our repository is to capture/represent all the research output of the institution- diversity is expected.

> Deposits, regardless of the type of resource, need to include sufficient supplementary information to preservation and future discovery, access and use.

The increased emphasis on the importance of high quality, descriptive metadata is an interesting result. The OAI-PMH metadata is based on the Dublin Core standard and was deliberately kept simple to provide a low entry threshold and encourage self-archiving. As the range of the different types of materials in repositories grows information about the items within the metadata must evolve to continue to be useful for discovery and for preservation.

Changes in scholarly publishing

A very large percentage, 82% of the respondents, believed that new types of electronic resources would change scholarly publishing. It is still not particularly clear though how exactly this is going to happen. One way seems to be by expanding the functionality of repositories and not just storing and providing access to electronic resources but by finding ways of connecting this information with other information systems.

> Many of our resources (images) are associated to preserved specimens in our biological collections, in such a way that we have not only an OAI repository but it also works as a image handling administrator for another information system.

From the interview data most repository managers also agreed that changes to what was considered scholarly publishing would somehow be altered by the presence of the institutional repository. However, there is little agreement on what types of changes will occur and how exactly this will happen. Repository managers appear to be balancing on one side a desire to include a wide range of material but with limited technical and practical experience on how to handle them, and on the other side, a need to collect formal research output, in particular e-prints.

Summary

The results from the interviews, the online survey and link analysis are presented in this chapter. Survey data was collected from repository managers in 31 different countries. The number of responses from each country is a reflection of the number of IRs per country.

The most popular accepted types of materials in repositories were ones that have paper-based equivalents. The same is true of the most frequent types of materials, with postprints being the most popular resource followed by other text-based documents. For non-formal resources a long and varied list of other types of materials that could be accepted within a repository was

produced. Most repository managers are well aware of the typological diversity of electronic resources that could potentially be deposited in a repository and we found three distinct approaches to dealing with them.

Decisions about what content could be deposited in the repository, was most commonly taken by a special committee and library staff, lecturers and researchers the most popular groups allowed to deposit. Repositories had different strategies for acquiring materials and differing collection policies depending on their repository drivers. With both self and mediated deposit we found backlogs of work, mainly due to lack of staffing. We also found discrepancies between collection policies and actual depositing practices most of them due to demands in the form of added on services, such as producing RAE returns, on the repository.

Monitoring usage statistics is quite common and over half the survey respondents believe that resources within the repository are frequently used. Although the potential of these statistics is noted, currently they are generally used for promotion and justification for funding. A target page and a source page typology for the link analysis was developed. Items linked to within the repositories are formal electronic publications followed by grey literature. Blogs are one of the frequent linkers to items within repositories.

Enhancing access to resources and preservation are the primary objectives of repositories and most are interested in acquiring a wide range of electronic resources. However, there was little agreement with regards to issues about quality control and ways of ensuring this. Some repositories continue to use peer-review and published as criterion for accepting materials. Others disagree and see this as a way of limiting, rather than enhancing repository capability.

In the following chapter the conclusions of the study are presented.

CHAPTER 5 –CONCLUSIONS

Introduction

In this chapter the main findings of the thesis are discussed in relation to the aims and objectives and conclusions presented. The main goal of the study was to determine the implications of repositories for the visibility and use of electronic resources by the academic community, in particular those resources that fall outside the framework of formal electronic publishing. Stated objectives were to uncover the types and distribution of non-formal electronic resources within repositories and to examine policies and manager's attitudes towards their value and importance, as well as exploring the extent to which these resources are presently used. This chapter also discusses the limitations of the study and the possibilities for further research in this area.

Repository manager's attitudes towards non-formal resources

Although formal electronic publications in the form of e-prints and books are, not surprisingly, the most popular type of material accepted for deposit in repositories, we also found that managers' attitudes towards non-formal electronic resources are generally quite positive and the large majority believe that repositories should contain a wide variety of resources. In particular, electronic grey literature (in the form of reports, working papers and theses) were the most frequent type of resources found within repositories after postprints. Additionally the survey results indicate that a broad range of materials are accepted for deposit in IRs as listed by the repository managers themselves. Although the origins of IRs stem from the need to provide infrastructure for disseminating pre prints, later to be replaced by the peer reviewed published postprint (ie the Harnard and Ginsparg model)[44], currently many IRs appear to incline more towards the Lynch model (Lynch 2003). Repositories are built to function as digital

[44] See *Summary of repository and IRs development* in Chapter 2

infrastructure for universities and to go beyond the sole acquisition of e-prints by managing and disseminating a broad range of digital output.

Although repository managers agreed in theory that a wide variety of material can be deposited, in practice we found that additional issues, in particular quality considerations, greatly affect the types of materials that are actually accepted for deposit. Repositories use external indicators, the main ones being published and/or peer-reviewed, to ensure a certain degree of quality of the materials. These indicators are associated with formal publications, such as journal articles or books and they do not apply generally to non-formal resources. So, although it would seem that the Lynch model would naturally lead IRs to manage a wide range of electronic resources, issues surrounding quality assurance and specifically the notion of published and/or peer reviewed material, greatly limit the types of materials handled to mainly formal publications. This policy has actually led some repositories to not accept preprints (which are not yet peer reviewed or published) at all. This is a surprising find, as it does not tie in with any of the three repository models (Ginsparg & Harnard, SPARC or Lynch[45]) described in the literature.

This tension between an all embracing attitude towards different types of resources and the desire to establish certain quality controls creates obstacles for the endorsement of non-formal resources as a valid form of scholarly communication and eventually possibly even publication. The SPARC position paper (Crow 2002) describes IRs as a means to revolutionize scholarly communication and publishing by providing certification of resources, a role traditionally undertaken by publishers. For non-formal resources, at least, we found no evidence that IRs were able to provide this function. Providing ways of certifying and validating new types of scholarly resources as well as alternative ways of demonstrating research excellence other than books and journals (Jaschik 2005) is currently under discussion within the academic world. However, committees are still grappling with how to actually do this and there is still no established criterion for evaluating and incorporating digital resources within academic

[45] See *Summary of repository and IRs development* in Chapter 2

recognition for tenure (MLA 2007). There is evidence that high quality digital materials are produced in departments that recognizes and supports this medium as a valid form of scholarship (Warwick, Galina et al. 2008).

We did find however, that although some IRs are attempting to uphold traditional notions of quality associated with published material, these intentions are particularly difficult to maintain in practice and we discovered evidence of blurring boundaries with regard to accepting or not a particular item type, such as working papers. In some cases we found IRs that welcomed only published/peer-reviewed materials were actually also accepting e-grey literature and using the university's name as an indicator of quality for readers. As mentioned by Kling and McKim (Kling and McKim 1999; Kling and McKim 2000) readers do not treat all material, peer-reviewed or otherwise, as equally trustworthy but rather rely on a series of processes and markers to make their assessment, such as the reputation of the authority responsible for producing the material. This suggests, in particular for grey literature, that IRs are utilizing the institution's good name and standing as a benchmark for quality and effectively. In doing so, it could be argued, taking on some of the traditional roles of publishers by collecting, ensuring quality and disseminating material (Tenopir and King 2001).

In the print world, one of the major concerns with grey literature is that it usually does not have a formal publisher, leading to limited distribution, access and retrieval (GreyNet 2004). Evidence from the link analysis suggests that IRs are boosting grey literature visibility. When we group together all grey literature categories, we find they account for the third most important target types below journal articles and book/book chapters. Of particular note is the fact that research projects source pages frequently link to working documents and reports within the IR. In the print environment, as grey literature, these types of documents would usually remain within the confines of a small research group and have limited dissemination. It could be supposed that research groups have found IRs to be a useful means to make their work more

readily available to their intended audience. The evidence thus points to IRs as important vehicles for the communication and distribution/publication of on-going research.

As mentioned in the introduction, IRs could lead to the emancipation of grey literature by providing an appropriate and convenient communication channel directly to interested parties. Consequently, it could be suggested that IRs may have had more impact on the distribution of grey literature than on formal scholarly publishing. If a characteristic of formal electronic publishing is that it is made public and delivered direct to target audiences, then if access and dissemination of electronic grey literature increases, the question is at what point does it stop being grey and become formal? IRs have given these types of materials a new and possibly very effective means of distribution/publishing making them more readily available and accessible than ever before. In this sense, institutional repositories are effectively blurring the distinction between distributed and published.

It is important to add however, that repositories have been particularly interested in grey literature because it is fairly easy to get hold of, there are fewer problems with copyright and in the particular case of mandated theses, have provided repositories with a steady flow of relatively trouble free material that is generally deposited by the students themselves. For example, almost all case study repositories had or were about to implement a thesis mandate. Managers were partially attracted to these types of materials because they were textual material that is easy to manage, especially in terms of copyright. This is different from books and journals where copyright has usually been transferred to the publishers.

We also found that acceptability of a resource for deposit was not always dictated from the top down through repository policies but on some occasions with fairly open repositories, its suitability was determined by the academic community itself. Once a resource is perceived as useful by the community then this may form part of the accepted resources. Thus, it could be argued that as more non-formal resources are deposited and validated by academic members,

new indicators that do not necessarily equate to peer review or published, will be developed. If IRs are providing endorsement of quality and availability, what other attributes are required for a source to be considered 'published' within this electronic environment scenario?

In order to address this question it is important to revisit Kling and McKim's differentiation between publishing as a communicative practice in which the key objective is to be read by the intended audience and publishing from a functionalist perspective where it serves to allocate authors status and reward (Kling and McKim 1999). It would seem that currently, despite the fact that the academic community recognizes the communicative value of non-formal resources, evaluating committees have still not incorporated these types of 'publications' into the reward system. Although the general discourse is about the future of research based on faster, better and different research capabilities such as for example, some of the main ideas of Cyberinfrastructure and eScience (Hey and Trefethen 2005) or reducing the unit of communication (Van de Sompel, Payette et al. 2004) so that electronic resources are not just print clones with hyperlinks (Odlyzko 2002; Henry 2003; Warner 2005); these types of new modes of scholarly publishing and communication are still not present in IRs. In spite of the interest in scholars handling, sharing and communicating through a wide variety of different types of resources as long as departments and institutions do not revise their promotion and tenure requirements, it is highly likely that these types of non-formal resources will continue to be considered unpublished. Institutions are still to find an effective mechanism that can evaluate and validate these resources.

Currently one of the main problems towards achieving this is that non-formal resources are still extremely heterogeneous and vary considerably in purpose and form. Articles and books are formats that have developed over a considerable period of time and publishers have set the standard and defined the norms of what is an acceptable research paper (Boyce 1999). Non-formal resources are varied and even among similar types (such as for example digitized critical annotations) there are no standards of presentation. Some newer resources, such as blogs or

wikis, are more standardized and easily identifiable but still not to the degree of the highly structured format of a formal publication, as in for example, a journal article. This is not necessarily desirable either, as flexibility allows for innovation and creativity. On the other hand, non standardized formats present great challenges for evaluation and measurement of impact and use. As new non-formal materials are created and deposited in repositories, it is therefore important to build a typology of resources. This is a first step towards understanding the impact that they are having and in the future provide us with criteria to recognize and value their importance in scholarly communication and publishing.

Typology of electronic resources

In the second chapter in the literature review, we identified a framework for defining electronic resources by dividing them into primary and secondary resources (Spark Jones, Bennett et al. 2005). Primary electronic resources were defined as those on which research is done as opposed to the latter through which primary resources are found and accessed. Examples of primary resources are articles or digitized manuscripts; and online bibliographies or online library catalogues are examples of secondary resources. For this study institutional repositories are a secondary resource, with the primary resources being the items deposited.

Although this served as a good starting point, primary resources is a broad category and in order to answer some of the main research questions, a more detailed approach was required. As defined initially, primary electronic resources encompass a wide variety of materials, such as digitized images, conference proceedings, books, graphics, satellite images, numerical data, text corpora and journal articles, to name but a few.

We found that IRs have not developed a particularly sophisticated typology or classification scheme for the types of materials they contain. Evidence from the interviews shows that most repository classification systems, in particular for non-formal resources, were developed rather

informally, some time ago and/or are still in process. When comparing repository classification systems, although there were some similarities, we found no consistency between naming schemes.

A main contribution of this study is the target page typology developed from the item types as defined by the repository. To develop this it was necessary to homogenize the different naming schemes in order to compare across repositories, including the formal and grey electronic publishing item types. The ensuing typology is quite general and granularity was compromised in favour of a broader scheme that allowed classification of all occurrences in the link analysis. In particular, grey literature items, such as working papers, research reports and other, were categorized differently across the case studies. In some cases the differences were slight but in others, it was necessary to merge conceptually different types into one category, like for example book and book chapter. This is a limitation of the typology, in particular if we wish to use it for usage or link analysis, as two conceptually different item types, book or book chapter are not differentiated. If repositories are to be an important source of more informal types of materials, the classification and identification of these types of resources will most likely need to be further developed. This also means that if in future we want to compare repository downloads and types of material, issues related to the variations in the classification schemes will need to be addressed.

This is particularly relevant for non-formal resources that are not text-based, such as video, images, audio, datasets and so forth. Most of these are currently classified under umbrella type categories such as multimedia, other and datasets, although the resources are heterogeneous and vary greatly in their purpose, content, format and intention. When attempting to study the impact of these resources on scholarly communication the fact that everything from satellite images to text corpora tends to be grouped together in one category is limiting and there is a danger of reaching general sweeping conclusions about their presence, distribution and use in IRs. It is important to attempt, at least, to look at how these resources are being understood and

managed by IRs in order to shed light on their impact and visibility, especially in the context of their present or future role in scholarly communication and publishing.

Developing a source page typology also proved a contentious but important issue. Certain genres are now more established than in previous classification schemes (Almind and Ingwersen 1997; Haas and Grams 1998; Crowston and Williams 2000) as for example with blogs. In addition academic web genres are relatively stable with source pages such as academic, departmental and research project pages being fairly consistent in content and structure and therefore easily identifiable. As new genres are developed and accepted it is likely that they will become more stable. This is probably true as well for non-formal item types within repositories. As certain resources become more common and better defined, their item types should be identified in the same way across different repositories.

Distribution and management of non-formal electronic resources

Repository managers indicated in the survey a broad range of material accepted within their repositories. However, we did not discover the expected breadth and scope of item types during the link analysis. Additional results from the survey show that the three least frequent types of items of resources are patents, datasets and software.

There could be a number of reasons for these results. In the case of the link analysis it is quite likely that the number of these types of resources available is relatively low and the small sample of links was not enough to find them. A larger sample could possibly have detected a larger variety. Additionally the repository classification systems as we have seen, are not particularly broad and some item types such as architectural drawings, digitized materials-XML encoded texts, art, raw and processed imagery obtained from satellite and aerial platforms, are all categorized under one item type, for example, image, obliterating the content and communicational differences.

Repository managers did indicate that non-formal resources are being produced and there is interest from some academics to make these available through the repository. However, we also found strong evidence to suggest that although repositories would be willing in theory to accept a wide range of material, in practice they do not actually know how to or want to handle these types of resources. This has led to repositories describing a much larger breadth of material than they actually manage.

In some cases, non-formal resources in particular are not necessarily deposited in IRs but rather in alternate specialized repositories. An example of this are datasets, in particular very large or dynamic ones that IRs can find difficult to handle. This has important implications if the IR is viewed as representative of the university's research output, as certain materials will be placed in external repositories. Many repository managers discussed the issue that one single repository would not fulfil all the university's needs. In this sense IRs are not viewed as a digital infrastructure that can handle all the university's digital output. One of the main points of IRs is that they will allow the university to exercise stewardship and ensure the preservation of these materials. If some of these materials are being placed in other repositories, then how can this be assured? There was no agreement on this point. Some managers felt that the objective of Open Access was to be available and it did not matter where whilst others expressed concern about the future of these materials. One of the case studies was a notable exception as they stressed preservation as the main driver for their repository and only accepted material that they actually stored. However, there was little interest in the dissemination or communicative aspects of the repository.

In some cases IRs do not use an external repository but rather they set up a separate repository to handle non-formal electronic resources, leaving the IR mainly for formal and grey electronic publishing. They are well aware of the interoperability challenges with regard to handling different institutional repositories and their relationship with subject or discipline repositories.

Even for those repositories that did handle non-formal resources we found that these were dealt with at a fairly basic level and that the nature of more complex or aggregated objects that combine a series of different documents types or are dynamic, are still not adequately managed under the current repository infrastructure. With the exception of some target items in one case repository, most of these tended to have a simple make up, i.e. one metadata item described one digital file in basic one to one relationship. This matter has been described and addressed by the OAI-ORE protocol[46] but this had been released only recently at the time of the interviews and was not discussed by any repository managers. The future of multiple genres linked or embedded to form a more complex pattern of communication (Crowston and Williams 2000) is still not clear within institutional repositories. Some managers mentioned the current limitations of the Dublin Core metadata schema and some had set up a separate repository specifically to utilize more complex metadata for multimedia, learning objects and other non-formal resources.

In particular during the interviews we found evidence to suggest that academics and indeed repository managers are pushing for these types of materials to be available in the repository. There is general awareness of materials mainly from the Arts departments and to some extent from the Humanities, that will not necessarily be text-based; such as images, digital art installations, photo exhibitions, music and so forth. All repository managers are very aware that these types of materials exist but they are still working on how best to deal with them. Quite a few are involved in projects that are specifically looking into handling, storing and disseminating a wide variety of electronic resources. So although currently there is no clear solution there is definitely recognition of the situation and work is in progress.

[46] See Chapter 2 *Towards new forms of electronic publishing*

Usage of non-formal resources

Usage of usage statistics

Usage statistics are becoming increasingly important and if access and dissemination are a key priority for a repository, usage statistics are a means of expressing and understanding the impact and usage of the resources contained within. We found evidence that almost all case study repositories are looking at usage statistics and consider them useful but they are not currently a priority. Most repository managers reported that at least until recently, the repository had been understaffed and the main priority had therefore thus far been to gather content for deposit. This meant that usage statistics were kept but analysis of them was limited. Currently most repositories do not differentiate their usage statistics by different types of materials, but as with their collection policy they tended to focus more on the use of journal articles. However, since the interviews were conducted both Eprints and DSpace have introduced download statistics as an easily installable option and this will most likely change the availability of download counts if not other types of usage statistics.

Repository managers perceived usage statistics to be particularly useful for proving the repository's impact to decision makers, funders and for advocacy. Interestingly enough no repository had altered their collection policy or changed their repository in any way as a result of information gathered by usage statistics. It seems from this that IRs are still working on proving their usefulness to the university authorities and usage statistics are primarily to justify their existence. It could be that in the future, if and when IRs have assured their long term survival, they may begin to focus more on using the statistics to understand their users' needs.

Interpreting and understanding usage statistics was an important issue for all repository managers. Many questioned what exactly could be deduced from the information that they were receiving, in particular downloads. This was true for all types of electronic resources within the repository. For example, what exactly do you measure and count? In terms of actual usage does

an abstract download mean the same and therefore count as an equivalent to an article download? Formal electronic publications tend to have a fairly standard structure. In the case of non-formal resources this is further complicated as we can encounter a variety of different types as well as structures and formats. How do we go about interpreting the download of an electronic resource if one is a three-minute audio and the other a three-hour video on a particular subject? What is to be done with more complex item types such as an audio of an interview with the accompanying word transcript? An important question with regard to non-formal resources is whether downloads as they are currently handled, can provide us with valuable information about their usage or do they require a more sophisticated and fine-tuned system?

Repository managers found a similar problem with determining how to count the number of items within a repository. This study strongly supports previous work (Westrienen van and Lynch 2005) that indicated the difficulties with using number of items as a means to describe or evaluate a repository, as some contain metadata only records whilst others store the digital object. If evaluative frameworks regarding IRs are going to take into consideration quantitative data such as number of items or usage downloads, it is important that these difficulties are kept in mind.

It is clear from the results that there is a lot more work to be done in this area. As with other aspects of electronic publishing, so far there has been more work done on usage statistics for formal and grey publishing (Jamali, Nicholas et al. 2005; Nicholas, Huntington et al. 2005b), and less so for non-formal resources. This is probably due to the fact that there is generally more interest in formal publishing but additionally non-formal resources present some particular challenges for measuring use. An important first step to resolving this is to work on genre classification as it is easier to count when you know what you are counting. Currently diverse resources are grouped together making it is difficult to accurately assess their usage. This is a similar situation to treating all formal electronic publishing as the same, when research (Kling

and McKim 1999) has insisted that disciplinary differences must be taken into account in order to better understand the phenomenon. A similar argument can be made for non-formal electronic resources.

Repository managers were quite disparaging about other more qualitative indicators of use of the resources within their repositories such as emails and other types of direct communication from users. These could be a useful source of information to find out exactly what people are using these resources for. However, there was more interest from the repository managers in having quantitative, metric based usage statistics and they treated this information as interesting but not particularly useful.

Using link analysis for determining usage

The link analysis study found that journal articles are the most frequently linked to items within repositories. Unfortunately due to the way that items were classified in the different repositories it was not possible to differentiate between links to preprints or to postprints. It seems that people prefer to cite print published versions of articles even if they consult the article online. It would be interesting to see if there is the same preference with people preferring to link to the postprint rather than the preprint. However, in order to do so the repository classification scheme must differentiate between pre and postprints and this is currently not the case for all repositories. The second most linked to electronic resources are books. Whether it be pre or postprints and books, there appears to be a strong preference for using traditional scholarly documents to link to. It could also be that there are more links to formal electronic publishing resources due to the fact that there are more of these types of materials deposited within the IRs. Although this is an area where more research is definitely required, it was considered out of the scope of the present study due to its focus on non-formal resources rather than formal publishing.

The source pages that linked the most frequently to items (excluding indexed/search engine and dot com linkpage) within repositories were blogs, academic homepages, discussion forums, news services and research project and as mentioned previously the most popular target types in repositories were articles, book/book chapters followed by conference proceedings. In contrast there were very few links found from formal electronic source pages such as e-journals, e-magazines or e-books to items within the repositories either formal or not. This suggests that digital genres with a print equivalent have migrated with similar characteristics to the online world and continue to link to more traditional sources.

There is a strong indication that formal traditional academic communication patterns continue to hold true in the electronic environment and for institutional repositories. Although technology would in theory allow these patterns to change, the actual practice takes longer. In the case of non-formal resources however, we did find evidence that indicate possible changes in patterns of communication. The link analysis data showed that in terms of source pages that link to repository items, the third most popular source type (eliminating dot com linkpages and search engines) were blogs, which is a digital only genre. Even more interestingly is that the links are not in lists (as in the case with publications) but are usually specifically arguing or exemplifying a point. The same holds true for discussion forums and mailing lists. These types of source pages could be considered more similar to a citation in terms of perceived usefulness than merely a reference to a publication. Although repositories may have not impacted traditional scholarly communication extensively, they could be having an important effect on more informal forms of communication such as the type found in blogs, discussions groups and mailing lists.

Link analysis was chosen as a method for examining the visibility and use of electronic resources within IRs. It had been used previously for studying research productivity and collaboration and academic relationships, invocation of authors and digital libraries (Cronin, Snyder et al. 1998; Thelwall 2003a; Thelwall 2003b; Payne and Thelwall 2004; Thelwall and

Harries 2004; Stuart, Thelwall et al. 2007; Zuccala, Thelwall et al. 2007) but never previously for IRs. Link analysis proved a useful methodology and has shown shed light on several undetected aspects about use of electronic resources. This study was purposely designed as a small, qualitative type study and more emphasis was placed on the use of the links rather than on the number of links. The relatively small size of the sample (552) allowed us to visit each target and source page and build a comprehensive and thorough target and source page typology. In particular, the target page typology which synthesizes item types from IRs is an important contribution of this study. This is because the lack of consistency of naming types was discovered as a particular difficulty for link analysis and other types of studies of non-formal resources. Defining units to be counted for link analysis is a problem shared by other link analysis studies (Thelwall 2002; Thelwall, Vaughan et al. 2005) and the target page typology paves the way for future more quantitative link analysis of IRs.

Impact on scholarly communication and publishing

Although it may seem fairly obvious we would expect a relationship between the main objectives of the repository and the types of materials that are collected within it. However, results show that for most institutional repositories achieving these objectives has been a challenge, not only because of the newness of the technology but also because objectives have been changing as repositories are being asked to fulfill more and different roles. As each added on service is built upon a repository, for example delivering the RAE or creating publication lists for academics homepages, there is an impact on the repository collection policy. This in turn will have a direct impact on what objectives the repository can actually fulfill. The result is that as new demands are made on repositories we encounter more contradictions between their objectives and their types of materials. An example of this are repositories that wanted to hold the digital object for all deposits but had to settle for metadata only records so that academics' complete publications lists could be generated from the repository.

Some IRs are finding that issues related to quality assurance in particular of the metadata and copyright clearance, together with a lack of dedicated staff for the repository, has created a back log in materials appearing in the repository. This challenges the notion that depositing an article is only a few keystrokes away. An academic may deposit an item and it can be months before the item actually appears in the repository. In some cases it seems that quality assurance and lack of staffing are significant barriers to rapid deposit in IRs. Another interesting area for further work for formal publications would be to investigate these time delays in more detail and see if they can be compared to publication times. This negative experience could possibly partially explain some academics reluctance to self-archive. More importantly for this study, some repositories will treat non-formal resources on a case-by-case basis, which probably implies an even lengthier waiting period. How will this affect the use of repositories for storing these types of materials? Although repository managers indicated estimated times, these varied widely depending on several factors (such as maternity leave, lack of funding, staff member on loan, changes in repository software and so forth) such that it is difficult to draw any conclusions about specific lengths of waiting time. The important thing to point out is that for IRs that have some sort of quality assurance procedure in place, deposits do not appear immediately in the repository.

The history and development of repositories has been closely related to the Open Access movement, although currently many institutional repositories are no longer primarily or solely OA driven. However, the collection policy focus continues to be on journal articles either as preprints or postprints. From the results there is little evidence to suggest that preprints have become important in the institutional repository world. The survey found that preprints are not one of the most frequent resources. This is backed up by the literature that discusses the difficulties in convincing academics to self-archive their work before it is published (Ashworth, Mackie et al. 2004; Jones, Andrew et al. 2006). Postprints nonetheless are quite frequent formal electronic resource types. However, it seems that post prints are deposited by repository staff rather than by authors themselves. This finding is supported by evidence from the interviews

with repository managers who describe the emphasis on acquiring journal articles for the repository. This model differs from the self-archiving one and as pointed out by several repository mangers themselves, the logistics of mediated deposited in particular for large universities is unworkable. A large number of repository staff would be needed if all the formal research output is to be covered in this manner.

A possible explanation for the emphasis on postprints, is that in the case of preprints, repository managers must rely on authors to self-archive, but when it comes to postprints they have been able to take a more proactive approach and find and deposit the articles themselves. Therefore the focus for repository collection has been mainly on published journal articles. It could be argued that this emphasis on postprints is due to the fact that repository managers prefer the peer-reviewed and published version of an article. What appears to be happening is that many institutional repositories are moving away from their initial objectives of providing access and dissemination to the university's research output as they find new added on services that can be provided through the institutional repositories. One example has been to use the institutional repository to deliver the RAE returns. This has meant that the final published version metadata is what is required for the repository. This has had two implications: one is that repositories with a full text only policy have had to accept metadata only records and the second is that more emphasis has been placed on the quality assurance of the metadata rather than on providing access to the digital object. For some this has meant moving away from the original OA objectives. More important for IRs that were set up as a tool for managing and administrating digital resources, is that they are also handling metadata only records. From the interviews we gathered evidence that suggests that some university authorities have been looking to use the IR as a complete registry of the university's formal electronic publishing output. It could be that, albeit surreptitiously, IRs are moving away from providing access towards providing a register for formal electronic publications.

The implications of this are that although in theory IRs are interested in either pre or postprints, the actual practice continues to uphold a more conventional scholarly publishing system by encouraging traditionally published articles and books. The way of measuring research productivity, in this case for the RAE, is still an important driver and document types are partially determined by how university research status is measured. In particular with repositories that incorporate metadata only records with links to the publishers' versions of the final article, it could be argued that they are now providing dissemination channels towards publisher websites. The separation of metadata and the digital object, an original concept of OAI-PMH to facilitate harvesting, has unintentionally allowed repositories of metadata-only or publications lists. Although some repository managers argue that the digital object will be made available as soon as it is possible, this is yet to be seen. Content type is determined to a great extent to valid forms of publication that are eventually used for evaluation rather than communication.

Repositories however, have had a great impact on grey literature and non-formal resources. For both they provide a channel for communication and dissemination that is revolutionary and unique. Repository managers and academics attitudes towards non-formal resources are generally quite positive, although ensuring quality and value is an important unresolved issue. Evaluation committees and other bodies that determine funding, tenure and so forth, must enter the discussion on alternative forms of providing valuable research output. This could possibly be the main key to future development, especially if academic communication and publishing is going to move away from the traditional print based media to more innovative forms of electronic publishing.

Limitations of the study

As pointed out in chapter two, the academic digital environment is still in development and changes occur rapidly. Consequently, this research has had to deal with a moving target in terms of new genres, repository development and cultural changes.

The typologies developed are general and will in all likelihood require updating in the near future. The repository scenario is also fast moving and over the course of the three years of this research, there have been constant modifications. One example is the changes to the OpenDOAR typology. Additionally new types of resources will appear and we should see an increase in certain types of materials, such as videos as these technologies become easier to handle. In this period the two main software platforms for institutional repositories, DSpace and EPrints, made changes that claim to make handling non-formal resources easier as well as making download statistics and other usage tools more accessible to implement and use.

The results showed that most repository managers believed that IRs were on the brink of changing their role within the university and becoming more incorporated and recognized within the system. One key issue is that most were managing to secure funding for a longer period and had recently hired one or more permanent members of staff. This could lead to big changes in IRs.

An increase in materials such as theses and posprints are expected in the immediate future as a result of mandates. We have already seen that mandates will naturally skew material frequency as for example, the large amounts of grey literature in the forms of theses in repositories. The distribution of material types will probably change.

It is therefore probable that some of the findings from this study will need to be revised in the light of technological and cultural changes that are currently in process. However, it is believed that the methodology employed will continue to be useful and additionally that the work done

can serve as an important benchmark for future studies in order to understand the history and development of managing and disseminating digital academic resources.

Further work

As mentioned in the previous section the research done is an important first step but there is still a lot of ground to cover in the future. Although we found little research on non-formal electronic resources it is clear from the results that this is an important issue for repositories. Many of the case studies repositories are currently involved in projects that are looking into how to handle datasets, cultural materials and other non-formal electronic resources. The findings from this study corroborate the initial proposal of this thesis about the importance of understanding these types of resources. This is an initial step towards a new and developing area where there is plenty of scope for continued work.

Now that the target page typology has been developed it would be interesting to undertake a larger, more quantitative type link analysis study that looks at a larger number of links and also more repositories. Both this study and previous work has suggested that there could be geographical differences with repository development, and a link analysis study could offer further insight into this issue.

Contributions of this study

This is the first study that looks at the variety and use of electronic resources within repositories with specific emphasis on identifying genre types. Previous studies have focused on formal electronic publications, specifically e-books and journals, but less attention has been paid to grey electronic publishing and non-formal electronic resources. This work argues that in order to better understand the impact of digital resources for scholarly communication and publishing,

we must pay more attention to the characteristics of different types of electronic resources to better assess their implications.

This study has therefore focused on understanding the perceived value of different genre types within repositories by surveying and interviewing repository managers. It has also attempted to define an initial typology of items within repositories. Previous studies (Almind and Ingwersen 1997; Crowston and Williams 2000; Rehm 2002; Thelwall 2002) have looked at genre identification but this is the first one of repository materials. This is complemented by a source page typology for pages that link to items within repositories, also an original contribution to the field.

Additionally although there is a large body of literature on the potential implications of repositories for scholarly communication and publishing there is less research on their actual effect. Some studies have focused on the impact of repositories for journal articles but once again little attention has been paid to other types of materials. As this study found, repository managers are interested mainly in access and dissemination of electronic resources as well as preservation issues. This study is the first to look specifically at usage of all types of resources. Additionally, this is the first link analysis study of institutional repositories and discovering the effectiveness of this methodology for understanding the use of electronic resources is a further contribution.

It is hoped that this in depth study will contribute to a better understanding of the nature and impact of different digital genres in academic environments.

Bibliography

Adair, J. R. (1997). TC: A Journal of Biblical Textual Criticism- A Modern Experiment in Studying the Ancients. *Journal of Electronic Publishing* **3**(1).

Aguillo, I. F., B. Branadino, et al. (2005). Posicionamiento en el web del sector académico iberoamericano. *Interciencia* **30**(12): 735-738.

AHRC - Arts and Humanities Research Council (2007). Announcement 14 May 2007. Published AHDS webpage 13 June 2007. Found at: Internet Archive: http://web.archive.org/web/20071012045201/ahds.ac.uk/news/futureAHDS.htm

Almind, T. and P. Ingwersen (1997). Informetric analyses on the World Wide Web: Methodological approaches to "webometrics". *Journal of Documentation* **53**(4): 404-426.

Andrew, T. (2003). Trends in Self-Posting of Research Material Online by Academic Staff. *Ariadne* (37).

Arms, W. (2002). Quality Control in Scholarly Publishing on the Web. *Journal of Electronic Publishing* **8**(1).

Arms, W. and R. Larsen (2007). *The Future of Scholarly Communication: Building the Infrastructure for Cyberscholarship*. NSF-JISC: 1-17.

Armstrong, C. J. and R. E. Lonsdale (2000). Scholarly monographs: why would I want to publish electronically? *The Electronic Library* **18**(1): 21-29.

Arnold, K. (1993). The Scholarly Monograph is Dead: Long Live the Scholarly Monograph. Scholarly Publishing on the Electronic Networks: The New Generation: Visions and Opportunities in Not-for-Profit Publishing: Proceedings of the Second Symposium, Washington, DC, Association of Research Libraries.

Ashworth, S., M. Mackie, et al. (2004). The DAEDALUS project, developing institutional repositories at Glasgow University: the story so far. *Library Review* **53**(5): 259-264.

Atkins, D. E. C., K. K. Droegemeier, et al. (2003). *Revolutionizing Science and Engineering Through Cyberinfrastructure*. Report of the National Science Foundation - Blue Ribbon Advisory Panel on Cyberinfrastructure, National Science Foundation: 84.

Auger, C. P. (1988). *Information Sources in Grey Literature*. London, England, Bowker-Saur.

Awre, C. and C. Baldwin (2006). *Focus on Access to Institutional Repositories*, Joint Information Systems Committee: 28.

Aymar, R. (2009). Scholarly communication in High-Energy Physics: past, present and future innovations. *European Review*. **17**(1): pp.33-51.

Bailey, C. J. W. (2005-2007). The Open Access Bibliography: Liberating Scholarly Literature with E-Prints and Open Access Journals. C. J. W. Bailey. 2007.

Bailey, C. J. W. (2006). Scholarly Electronic Publishing Bibliography.

Bailey, C. J. W., K. Coombs, et al. (2006). *Institutional Repositories - SPEC Kit 292*. SPEC Kits. L. A. George, ARL: 13-21.

Banks, M. (2005). Towards a Continuum of Scholarship: The Eventual Collapse of the Distinction Between Grey and non-Grey Literature. Open Access to Grey Resources, Nancy, France.

Bar-Ilan, J. (2005). What do we know about links and linking? A framework for studying links in academic environments. *Information Processing and Management* **41**(3): 973-986.

Barkier, J.G. (2008). *Perceptions of Developing Trends in Repositories*. Survey Results for the SPARC Digital Repositories Meeting 2008.

Beck, S. E. and K. Manuel (2004). *Practical Research Methods for Librarians and Information Professionals*. New York, Neal-Schuman Publishers.

Björneborn, L. and P. Ingwersen (2004). Toward a Basic Framework for Webometrics. *Journal of the American Society for Information Science and Technology* 55(14): 1216-1227.

Bollen, J. and H. Van de Sompel (2008a). MESUR: implications of usage-based evaluations of scholarly status for open repositories. Third International Conference on Open Repositories 2008, Southampton.

Bollen, J. and H. Van de Sompel (2008b). Usage Impact Factor: the effects of sample characteristics on usage-based impact metrics. *Journal of the American Society for Information Science and Technology* **59**(1): 1-14pp.

Bouma, G. D. (2000). *The Research Process*. Australia, Oxford University Press.

Boyce, P. B. (1999). We Are All Publishers Now. What Does That Mean? Proceedings from the Internet Society. San Jose California, USA.

Brown, P. O., D. Cabell, et al. (2003). Bethesda Statement on Open Access Publishing. http://www.earlham.edu/~peters/fos/bethesda.htm (visited 17/02/2009)

Bullinger, H.-J., K. M. Einhäupl, et al. (2003). Berlin Declaration on Open Access to Knowledge in the Sciences and the Humanities. http://oa.mpg.de/openaccess-berlin/berlindeclaration.html (visited 13/11/2008).

Burke, L. A. and K. E. James (2006). Using online surveys for primary research data collection: lessons from the field. *International Journal of Innovation and Learning* **3**(1).

Cameron, M. (2005). Why People Don't Read Online and What to do About it. *Ubiquity* **6**(40).

CENLFEP Committee (2005). Statement of the Development and Establishment of Voluntary Deposit Schemes For Electronic Publications. Annual Conference of European National Libraries, Luxembourg.

Chan, L. (2004). Supporting and Enhancing Scholarship in the Digital Age: The Role of Open-Access Institutional Repositories. *Canadian Journal of Communication* **29**: 277-300.

Chan, L., D. Cuplinskas, et al. (2002). Budapest Open Access Initiative. http://www.soros.org/openaccess/read.shtml (visited 30/04/2009)

Charmaz, K. (2006). *Constructing Grounded Theory- A Practical Guide Through Qualitative Analysis*. London, SAGE.

Cronin, B. (2003). Scholarly Communication and Epistemic Cultures. *New Review of Academic Librarianship* **9**(1): 1-24

Cronin, B. (2008). The sociological turn in information science. *Journal of Information Science* **34**(4): 465-475.

Cronin, B., H. W. Snyder, et al. (1998). Invoked on the Web. *Journal of the American Society for Information Science* **49**(14): 1319-1328.

Crow, R. (2002). *The Case for Institutional Repositories: A SPARC Position Paper*. Washington DC, Scholarly Publishing and Academic Resources Coalition: 37.

Crowston, K. and M. Williams (2000). Reproduced and emergent genres of communication on the world wide web. *The Information Society* 16: 201-215.

Culter, D. (1999). Grey Literature in Energy: A shifting paradigm. New Frontiers in Grey Literature, GL '99, GreyNet. Washington D.C., USA.

Davila, J. A., L. A. Núñez, et al. (2006). www.saber.ula.ve: Un ejemplo de repositorio institucional universitario. *Interciencia* 31(001): 29-36.

Davis, P. M. and M. J. L. Connolly (2007). Institutional Repositories - Evaluating the Reasons for Non-use of Cornell University's Installation of DSpace. *D-Lib Magazine* 13(3/4).

Day, C. (1997). The Specialized Scholarly Monograph in Crisis or How Can I Get Tenure If You Won't Publish My Book? Digital Alternatives: Solving the Problem or Shifting the Costs? Association of Research Libraries Resources. http://www.arl.org/resources/pubs/specscholmono/day~print.shtml (visited 14/04/2009)

Dey, I. (2004). "Grounded Theory" in C. Seale, G. Gobo, et al. (Eds), *Qualitative Research Practice*, SAGE.

Dunning, A. (2006). The Tasks of the AHDS: Ten Years On. *Adriadne* 48.

Duranceau, E. F. (2008). The "Wealth of Networks" and Institutional Repositories: MIT, DSpace, and the Future of the Scholarly Commons. *Library Trends* **57**(2): pp. 244-261.

Emly, M. (2007). *MIDESS Project Final Report*, JISC & CURL: 23pp.

Eysenbach, G. (2006). Citation Advantage of Open Access Articles. *PLoS Biol* **4**(5).

Fillmore, L. (1993). Online Publishing: Threat or Menace?, Graphics Communications Association Online Publishing Conference. Pittsburg, USA.

Flores Cuesta, G. and N. Sánchez Tarragó (2007). Los repositorios institucionales: análisis de la situación internacional y principios generales para Cuba. *Acimed* **16**(6).

Frankel, M. S., R. Elliot, et al. (2000). Defining and Certifying Electronic Publication in Science: A Proposal to the International Association of STM Publishers. *Learned Publishing* **13**(4): 251-258.

Gadd, E., C. Oppenheim, et al. (2003a). The Intellectual Property Rights Issues Facing Self-Archiving: Key finding of the RoMEO project. *D-Lib Magazine* **9**(9).

Gadd, E., C. Oppenheim, et al. (2003b). The RoMEO Project: Protecting metadata in an open access environment. *Adriadne*(36).

Gadd, E., C. Oppenheim, et al. (2003c). RoMEO studies 2: How Academics Want to Protect their Open-Access Research Papers. *Journal of Information Science* **29**(5): 333-356.

Gerring, J. (2007). *Case Study Research - Principles and Practices*. Cambridge University Press.

Ginsparg, P. (1996). Winners and Losers in the Global Research Village, Scientist's View of Electronic Publishing and Issues Raised, UNESCO Conference. Paris, France.

Gorman, G. E. and P. Clayton (2005). *Qualitative Research for the Information Professional*. London, Facet Publishing.

Greenstein, D. and J. Trant (1996). AHDS: Arts and Humanities Data Service. *Adriadne*(4).

Grenquist, P. (1997). Why I Don't Read Electronic Journals: An Iconoclast Speaks Out. *Journal of Electronic Publishing,* **3**(1).

GreyNet (2004). Grey Literature Network Service. http://www.greynet.org/ (visited 17/11/2007)

Guédon, J.-C. (2001a). Beyond Core Journals and Licenses: The Paths to Reform Scientific Publishing. *ARL Bimonthy Report* 218.

Guédon, J.-C. (2001b). In Oldenburg's Long Shadow: Libraries, Research Scientists, Publishers, and the Control of Scientific Publishing. Association of Research Libraries, Membership Meeting Proceedings.

Haas, S. W. and E. S. Grams (1998). Page link and classifications: Connecting diverse resources. Proceedings of Digital Libraries - Third ACM conference on Digital Libraries. Houston, USA.

Hagedorn, K. and J. Santelli (2008). Google Still Not Indexing Hidden Web URLs. *D-Lib Magazine* **14**(7/8).

Harnard, S. (1990). Scholarly Skywriting and the Prepublication Continuum of Science Inquiry. *Psychological Science* 1: 342-343.

Harnard, S. (2001). The Self-Archiving Initiative. *Nature* 410: 1024-1025.

Harnard, S. and T. Brody (2004). Comparing the Impact of Open Access (OA) vs. Non-OA Articles in the Same Journals. *D-Lib Magazine* **10(**6).

Harnard, S., L. Carr, et al. (2003). Mandated Online RAE CVs linked to university eprint archives: Enhancing UK research impact and assessment. *Adriadne*(35).

Harnard, S., T. Brody, et al. (2004). The Access/Impact Problem and the Green and Gold Roads to Open Access. *Serials Review* 30(4): 310-314.

Hart, C. (2001). *Doing a Literature Search - A Comprehensive Guide for the Social Sciences*. Oxford: Sage Publications.

Harvey, R. (2005). “Introduction” in K. Williamson (Ed), *Research Methods for students, academics and professionals - Information management and systems*. New South Wales: Centre for Information Studies - Charles Sturt University.

Hecht, F., B. K. Hecht, et al. (1998). The Journal "Impact Factor: A Misnamed, Misleading, Misused Measure." *Cancer, Genetics and Cytogenetics* 104(2): 77-81.

Heery, R. and S. Anderson (2005). *Digital Repositories Review*. UKOLN and AHDS: 33.

Henry, G. (2003). On-line Publishing in the 21st Century: Challenges and Opportunities. *D-Lib Magazine* **9**(10).

Hey, T. and A. E. Trefethen (2005). Cyberinfrastructure for e-Science. *Science* **308**(5723): 817-821.

Holmström, J. (2004). Calculating the Cost per Article of Reading Open Access Articles. *D-Lib Magazine* **10**(1).

Houghton, J. W. (2001). Crisis and transition: the economics of scholarly communication. *Learned Publishing* 14: 167-176.

Houghton, J. W., C. Steele, et al. (2004). Research practices and scholarly communication in the digital environment. *Learned Publishing* **17**(3): 231-249.

Hubbard, B. (2003). SHERPA and Institutional repositories. *Serials* **16**(3): 243-247.

Humphreys, S. R. (1997). Why Do We Write Stuff That Even Our Colleagues Don't Want To Read? The Specialized Scholarly Monograph in Crisis or How Can I Get Tenure If You Won't Publish My Book? Association of Research Libraries, Resources. http://www.arl.org/resources/pubs/specscholmono/Humphreys.shtml (visited 29/01/2009)

Huntington, P., D. Nicholas, et al. (2006). Obtaining subject data from log files using deep log analysis: case study OhioLINK. *Journal of Information Science* **32**(4).

Jamali, J. H., D. Nicholas, et al. (2005). The use and users of scholarly e-journals: a review of log analysis. *Aslib Proceedings* **57**(6): 554-571.

Jaschik, S. (2005). Radical Change for Tenure. *Inside Higher Ed. News*. http://www.insidehighered.com/news/2005/12/30/tenure (visited 28/09/2008)

JISC (2005). *Digitisation in the UK: The case for a UK framework*. JISC and Loughborough University: 36.

Johnson, R. K. (2002). Institutional Repositories Partnering with Faculty to Enhance Scholarly Communication. *D-Lib Magazine* **8**(11).

Jones, D. E. (1998). From Language Barriers to Contemporaneous Minds. *Journal of Electronic Publishing* **3**(3).

Jones, R., T. Andrew, et al. (2006). *The Institutional Repository*. Chandos Publishing.

Joseph, H. (2006). The Scholarly Publishing and Academic Resources Coalition: An evolving agenda (A SPARC Article). *C&RL News* **67**(2).

Kim, H. J. (2000). Motivations for Hyperlinking in Scholarly Electronic Articles: A Qualitative Study. *Journal of the American Society for Information Science and Technology* **51**(10): 887-889.

King, D. W. and C. Tenopir (1999). Evolving journal costs: implications for publishers, libraries, and readers. *Learned Publishing* 12: 251-258.

Kircz, J. G. (2001). New practices for electronic publishing. 1: Will the scientific paper keep its form? *Learned Publishing* **14**(4): 265-272.

Kircz, J. G. (2002). New practices for electronic publishing 2: New forms of the scientific paper *Learned Publishing* **15**(1): 27-32.

Kircz, J. G. (2005). *Institutional Repositories, a new platform in Higher Education and Research*. SURF-DARE, KRA Publishing Research.

Kling, R. (1999). What is Social Informatics and Why Does it Matter? *D-Lib Magazine* 5(1).

Kling, R. (2000). Learning About Information Technologies and Social Change: The Contribution of Social Informatics. *The Information Society* 16: 217-232.

Kling, R. and G. McKim (1999). Scholarly Communication and the Continuum of Electronic Publishing. *Journal of the American Society for Information Science* 50(10): 890-906.

Kling, R. and G. McKim (2000). Not just a matter of time: Field differences and the shaping of electronic media in supporting scientific communication. *Journal of the American Society for Information Science* 51(14): 1306-1320.

Kling, R., H. Rosenbaum, et al. (2005). *Understanding and Communicating Social Informatics: A Framework for Studying and Teaching the Human Contexts of Information and Communication Technologies*: Information Today, Inc.

Lambert, S., B. Matthews, et al. (2005). Grey literature, IRs, and the organizational context. 7th International Conference on Grey Literature: Open Access to Grey Resources, Nancy, France.

Lawrence, S. (2001). Free online availability substantially increases a paper's impact. *Nature* **411**(6837): 521.

Levy, Y. and T. J. Ellis (2006). A Systems Approach to Conduct an Effective Literature Review in Support of Information Systems Research. *Informing Science Journal* 9: 181-212.

Liu, Z. (2005). Reading behavior in the digital environment. *Journal of Documentation* **61**(6): 700-712.

Lynch, C. (2001). The Battle to Define the Future of the Book in the Digital World. *First Monday* **6**(6).

Lynch, C. (2003). Institutional Repositories: Essential Infrastructure for Scholarship in the Digital Age. ARL Bimonthy Report 226.

Lynch, C. and J. K. Lippincott (2005). Institutional Repository deployment in the United States as of early 2005. *D-Lib Magazine* 11(9).

Mansourian, Y. (2006). Adoption of grounded theory in LIS research. *New Library World* **107**(1228/1229): 386-402.

McCook, A. (2006). Is Peer Review Broken? *The Scientist*. 20: 26-30.

McCown, F., X. Liu, et al. (2005). Search Engine Coverage of the OAI-PMH Corpus, IEEE Internet Computing, 10(2):66-73.

McDowell, C. S. (2007). Evaluating Institutional Repository Deployment in American Academe Since Early 2005 - Repositories by Numbers, Part 2. *D-Lib Magazine* **13**(9/10).

McKiernan, G. (2000). arXiv.org: the Los Alamos National Laboratory e-print server - New products in grey literature. *The International Journal on Grey Literature* **1**(3):127-138.

McMillan, G. (1999). Perspectives on Electronic Theses and Dissertations. New Frontiers in Grey Literature, GL '99, Washington, D.C. USA.

Meadows, A. J. (1998). *Communicating Research*, Academic Press.

Meckseper, C. and C. Warwick (2003). The Publication of Archaeological Excavation Reports Using XML. *Literary and Linguistic Computing: Journal of the Association for Literary and Linguistic Computing* **18**(1): 63-75.

Mele, S. (2009). Open Access Publishing in High-Energy Physics: the SCOAP3 model. *OCLC Syst. Serv.* **25**(1):20-34.

Millington, P. (2008) OpenDOAR and ROAR. RSP Services Day, Nottingham, http://www.opendoar.org/documents/RSPservices2007-04-23.ppt (visited 01/07/09).

Muller M., H. Ulrich, et al. (2008). Repository Case Histories. OR2008 Third International Conference on Open Repositories, University of Southampton. Southampton, UK.

MLA (2007). *Report of the MLA Task Force on Evaluating Scholarship for Tenure and Promotion*, Modern Language Association of America.

Nelson, M. and K. Maly (1999). Preserving the Pyramid of STI Using Buckets. New Frontiers in Grey Literature, GL '99. Washington D.C., USA.

Nicholas, D., P. Huntington, et al. (2005a). Open access journal publishing: the views of some of the world's senior authors. *Journal of Documentation* **61**(4): 497-519.

Nicholas, D., P. Huntington, et al. (2005b). Scholarly journal usage: the results of deep log analysis. *Journal of Documentation* **61**(2): 248-280.

Nicholas, D., P. Huntington, et al. (2007). E-books: how are users responding? *Update* **6**(11): 29-31.

Nicholas, D., I. Rowlands, et al. (2008). UK scholarly e-book usage: a landmark survey. *Aslib Proceedings* **60**(4): 311-334.

NSF (2006). *Reports and Workshops Relating to Cyberinfrastructure and Its Impacts*. National Science Foundation.

Odlyzko, A. (1995). Tragic Loss or Good Riddance? The Impending Demise of Traditional Scholarly Journals. *International Journal of Human-Computer Interaction* **42**(1): 71-122.

Odlyzko, A. (2002). The rapid evolution of scholarly communication. *Learned Publishing* **15**(1): 7-19

Okerson, A. and J. O'Donnell, Eds. (1995). *Scholarly Journals at the Crossroads: A Subversive Proposal for Electronic Publishing, An Internet Discussion about Scientific Journals and Their Future.* Washington, Association of Research Libraries. http://www.arl.org/sc/subversive/index.shtml (visited 18/04/2008)

Paskin, N. (2003). DOI- A 2003 Progress Report. *D-Lib Magazine* **9**(6).

Payne, N. and M. Thelwall (2004). A statistical analysis of UK academic web links. *Cybermetrics* **8**(1).

Peters, J. (1996). The Hundred Years War Started Today: An exploration of electronic peer review. *Journal of Electronic Publishing* **1**(1/2).

Pinfield, S. (2003). Open Archives and UK Institutions. *D-Lib Magazine* **9**(3).

Pinfield, S. (2005). A mandate to self-archive? The role of open access institutional repositories. *Serials* **18**(1):30-34.

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

Punch, K. F. (2005). *Introduction to Social Research - Qualitative and Qualitative Approaches.* London:Sage.

Ramaiah, C. K. (2005). An overview of electronic books: a bibliography. *The Electronic Library* **23**(1):17-44.

Ramalho, A. M. and M. d. Castro Neto (2002). The role of eprint archives in the access to, and dissemination of, scientific grey literature: LIZA- a case study by the National Library of Portugal. *Journal of Information Science* **28**(3):231-241.

Ramirez, E. (2003). The impact of the internet on the reading practices of a university community: the case of the UNAM. Proceedings of the 69th IFLA General Conference and Council. Berlin, Germany.

Rapley, T. (2004). "Interviews" in C. Seale, G. Gobo, et al. (Eds.), *Qualitative Research Practic*e (pp.15-33). SAGE:.

Rehm, G. (2002). Towards automatic web genre identification. Hawaii International Conference on System Sciences, IEEE Computer Society. Hawaii, USA.

Rieh, S. Y., K. Markey, et al. (2007). Census of Institutional Repositories in the U.S. A comparison across institutions at different stages of IR development. *D-Lib Magazine* **13**(11/12).

Robbin, A. and R. Day (2006). *On Rob Kling: The Theoretical, the Methodological, and the Critical.* Boston: Springer.

Rowlands, I. and D. Nicholas (2005). *New Journal Publishing Models: An international survey of senior researchers,* CIBER.

Rowlands, I., D. Nicholas, et al. (2007). What do faculty and students really think about e-books? *Aslib Proceedings* **59**(6):489-511.

Royster, P. (2007). Publishing Original Content in an Institutional Repository. *Serials Review* **34**(1):pp.27-30.

Rubin, H. J. and I. S. Rubin (2005). *Qualitative Interviewing - The Art of Hearing Data*, Sage.

Sale, A. (2005). *The key things to know, University of Tasmania - ePrints repository*. Unpublished report, UTAS ePrints Repository. http://eprints.utas.edu.au/223/ (visited 10/05/2008)

Sale, A. (2006). The acquisition of open access research articles." *First Monday* **11**(9).

Sale, A. (2007). The patchwork mandate. *D-Lib Magazine* **13**(1/2).

Salo, D. (2008). Innkeeper at the Roach Motel. *Library Trends* **57**(2):98-123.

Seglen, P. O. (1997). Why the impact factor of journals should not be used for evaluating research. *BMJ* **314**(7079):497.

Shreeves, S. L. and M. H. Cragin (2008). Introduction: Institutional Repositories: Current State and Future. *Library Trends* **57**(2):89-97.

Silverman, D. (2005). *Doing Qualitative Research*. London:Sage.

Smith, M., M. Barton, et al. (2003). DSpace - An Open Source Dynamic Digital Repository. *D-Lib Magazine* **9**(1).

Spark Jones, K., R. Bennett, et al. (2005). *E-resources for research in the humanities and social sciences: A British Academic Policy Review*, British Academy.

Stuart, D., M. Thelwall, et al. (2007). UK academic web links and collaboration – an exploratory study. *Journal of Information Science* **33**(2):231-246.

Suber, P. (2004). *Open Access Overview*. http://www.earlham.edu/~peters/fos/overview.htm (visited 3/05/2008)

Swan, A. and S. Brown (2003). Authors and Electronic Publishing: What Authors Want from the New Technology. *Learned Publishing* **16**(1): 28-33.

Tanner, K. (2002). "Survey research". In K. Williamson (Ed), *Research methods for students, academics and professionals - 'Information management and systems.*. New South Wales: Centre for Information Studies Charles Sturt University: 89-110.

Tansley, R. and S. Harnard (2000). Eprints.org Software for Creating Institutional and Individual Open Archives. *D-Lib Magazine* **6**(10).

Tashakkori, A. and C. Teddlie (1998). *Mixed Methodology: Combining Qualitative and Quantitative Approaches.* California: Sage.

Tenopir, C. and D. W. King (2001). Lessons For the Future of Journals. *Nature Web Debates*. http://www.nature.com/nature/debates/e-access/Articles/tenopir.html (visited 04/12/2008)

Thatcher, S. (2005) "From Dissertation to Book- Comment". In Scott Jaschik., Radical Change for Tenure, News Inside Higher Ed. http://www.insidehighered.com/news/2005/12/30/tenure (visited 16/04/2009)

Thelwall, M. (2002). Conceptualizing documentation on the web: an evaluation of different heuristic-based models for counting links between university web sites. *Journal of the American Society for Information Science and Technology* **53**(12): 995-1005.

Thelwall, M. (2003a). Web use and peer interconnectivity metrics for academic web sites. *Journal of Information Science* **29**(1):1-10.

Thelwall, M. (2003b). What is this link doing here? Beginning a fine-grained process of identifying reasons for academic hyperlink creation. *Information Research* **8**(3).

Thelwall, M. (2008). How are Social Network Sites Embedded in the Web? An Exploratory Link Analysis. *Cybermetrics* 12(1).

Thelwall, M. (2008). How are social network sites embedded in the web? An exploratory link analysis. *International Journal of Scientometrics, Informetrics and Bibliometrics* **12**(1).

Thelwall, M. and G. Harries (2003). The Connection Between the Research of a University and Counts of Links to Its Web Pages: An Investigation Based upon a Classification of the Relationships of Pages to the Research of a Host University. *Journal of the American Society for Information Science and Technology* **54**(7): 594-602.

Thelwall, M. and G. Harries (2004). Do Web Sites of Higher Rated Scholars Have Significantly More Online Impact? *Journal of the American Society for Information Science and Technology* **55**(2): 149-159.

Thelwall, M., L. Vaughan, et al. (2005). Webometrics. *Annual Review of Information Science and Technology* 39.

Thomas, C. and R. H. McDonald (2007). Measuring and Comparing Participation Patterns in Digital Repositories - Repositories by the Numbers, Part 1. *D-Lib Magazine* **13**(9/10).

Tuttle, M. (1989). Editor's Introduction to the Newsletter on Serials Pricing Issues. *Newsletter on Serials Pricing Issues* **1**(2`).

Unsworth, J. (1997). Documenting the Revinvention of Text: The Importance of Failure. *Journal of Electronic Publishing* **3**(2).

Unsworth, J. (2005). *Pubrarians and Liblishers: New Roles for Old Foes*. Unpublished. Keynote address Annual Meeting of the Society for Scholarly Publishing. Boston, USA.

Van de Sompel, H. and C. Lagoze (2000). The Santa Fe Convention of the Open Archives Initiative. *D-Lib Magazine* **6**(2).

Van de Sompel, H., S. Payette, et al. (2004). Rethinking Scholarly Communication: Building the System that Scholars Deserve. *D-Lib Magazine* **10**(9).

Van der Kuil, A. and M. Feijen (2004). The Dawning of the Dutch Network of Digital Academic Repositories (DARE): A Shared Experience. *Adriadne*(41).

Vickery, B. C. (2000). *Scientific Communication in History*. Maryland:Scarecrow Press, Inc.

Vine, R. (2006). Google Scholar. *Journal of Medical Library Association* **94**(1):97-99.

Ware, M. (2004a). Institutional repositories and scholarly publishing. *Learned Publishing* **17**(2): 115-124.

Ware, M. (2004b). *Pathfinder Research on Web-based Repositories*. Publisher & Library/Learning Solutions (PALS).

Warner, S. (2005). The transformation of scholarly communication. *Learned Publishing* **18**(3): 177-185.

Warwick, C. (2002). "Electronic publishing: what difference does it make?". In S. Hornby and Z. Clark (Eds), *Challenges and Change in the Information Society.* London: Facet Publishing: 200-218.

Warwick, C., I. Galina, et al. (2008). The master builders: LAIRAH research on good practice in the construction of digital humanities projects. *Literary and Linguistic Computing: Journal of the Association for Literary and Linguistic Computing* 23(3): 383-396.

Warwick, C., M. Terras, et al. (2006). *The LAIRAH Project: Log Analysis of Digital Resources in the Arts and the Humanities Final Report to the Arts and Humanities Research Council*. London, Arts and Humanities Research Council: 60pp.

Warwick, C., M. Terras, et al. (2008). If You Build It They Will Come? The LAIRAH Study: Quantifying the Use of Online Resources in the Arts and the Humanities through Statistical Analysis of User Log Data. *Literary and Linguistic Computing: Journal of the Association for Literary and Linguistic Computing* **23**(1): 85-102.

Westrienen van, G. and C. Lynch (2005). Academic Institutional Repositories: Deployment Status in 13 Nations as of Mid 2005. *D-Lib Magazine* **11**(9).

Wheary, J. and B. F. Schutz (1997). Living Reviews in Relativity: Making an Electronic Journal Live. *Journal of Electronic Publishing* **3**(1).

Wilkinson, D., G. Harries, et al. (2003). Motivations for academic web site interlinking: evidence for the Web as a novel source of information on informal scholarly communication. *Journal of Information Science* **29**(1): 49-56.

Williamson, K. (2002). "*Research methods for students, academics and professionals- Information management and systems*". In K. Williamson (Ed), *Research Methods for students, academics and professionals*. New South Wales: Centre for Information Studies - Charles Sturt University.

Willinsky, J. (2003). The Nine Flavours of Open Access Scholarly Publishing. *Journal of Postgraduate Medicine* **49**(3): 263-267.

Willis, G. (1996). Embracing electronic publishing. *Internet Research: Electronic Networking Applications and Policy* **6**(4): 77-90.

Wilson, T. (1981). Sociological aspects of information science. *International Forum of Information and Documentation* **6**(2): 13-18.

Zuber, P. A. (2008). A Study of Institutional Repository Holdings by Academic Discipline. *D-Lib Magazine* **14**(11/12).

Zuccala, A., M. Thelwall, et al. (2007). Web intelligence analyses of digital libraries - A case study of the National electronic Library for Health (NeLH). *Journal of Documentation* **63**(4): 558-589.

Appendices

Annex 1: Online survey for repository managers

I. INTRODUCTION

Questionnaire for Repository Managers

This survey is aimed at gathering information from institutional repository managers about the types of materials that are collected within your repository and depositing behaviour. All replies are confidential and will only be used in combination with those of other participants.

The survey is part of a project on the use of electronic resources in institutional repositories, undertaken at CIBER (Centre for Information Behaviour and the Evaluation of Research), University College London. For more detailed information on project please visit: http://www.ucl.ac.uk/~uczciga/repositories/

The survey should take approximately 15 min to complete and your replies will be much appreciated.

Please email i.russell@ucl.ac.uk if you have any problems technical difficulties with this survey.

II. REPOSITORY INFORMATION

Please provide information about the repository you currently run. The survey is designed for gathering information about one repository, so if you manage more than one repository please select the repository with the most records or if possible, fill in one survey per repository.

1. Repository name:

2. URL of repository: http://

3. Name or university or organization hosting the repository:

4. Country of hosting university:

5. How would you classify your repository in terms of development? [Please select one option]

Prototype
Recently launched/Initial stage
Fully operational repository
I don't know
Other (please specify)

6. How old is the repository you currently administer? [Please select one option]

Less than 1 year
1-2 years
2-3 years
4-5 years
more than 5 years

7. How many items are currently in your repository? [Please select one option]

0-100
101-500
501-800
801-1000
1001-5000
5001-8000
8001- 10,000
10,001-20,000
20,001-50,000
50,001-80,000
80,001-100,000
over 100,000

III. REPOSITORY MATERIALS

Please provide information about the types of electronic resources that your repository stores

8. Please select the type of electronic resources that are accepted within your repository. [Tick as many boxes as apply]

	Peer reviewed	Non-peer reviewed	Don't know
Books and book chapters			
Conference proceedings			
Workshop papers			
Datasets and databases			
Learning objects			
Audio-visual materials and multimedia			
Patents			
Postprints			
Preprints			
References/bibliographies			
Software			
Theses and dissertations			
Reports			
Working papers			
Images, maps, diagrams			
Administrative documents			

9. What other types of electronic resources are deposited which are not mentioned above? Please indicate if these are peer-reviewed.

10. What are the 3 most and least frequent content types in your repository? [Tick a maximum of three boxes for each]

	Most frequent	Less frequent
Books and book chapters		
Conference proceedings		
Workshop papers		
Datasets and databases		
Learning objects		
Audio-visual materials and multimedia		
Patents		
Postprints		
Preprints		
References/bibliographies		
Software		
Theses and dissertations		
Reports		
Working papers		
Images, maps, diagrams		
Administrative documents		

IV. REPOSITORY DEPOSITS

Please provide information about what types of resources are deposited in your repository and who deposits them.

11. Who decides on which type of electronic resources can be deposited in the repository? <Tick as many boxes as apply>

Myself
Library
Special committee
Computing Services
Don't know
Other (please specify)

12. What user group(s) are authorized to deposit materials in the repository? [Tick as many boxes as apply]

Lecturers/Researchers
Library staff
Administrative Assistants
Students
Other (please specify)
I don't know (please go to next section)

V. REPOSITORY DEPOSITS (cont.)

13. A user may deposit material within a repository as an AUTHOR (an item authored by them) or an item authored by someone else (on behalf of a THIRD PARTY).
In your experience how actively do these groups deposit their own work (as AUTHORS and not on behalf of a third party)?

	Extremely active	Frequently active	Occasionally active	Rarely or never active	Not authorized to deposit
Lecturers/Researchers					
Library staff					
Administrative Assistants					
Students					
Other (please specify)					

14. How actively do these groups deposit in general (as authors and/or on behalf of a third party)

	Extremely active	Frequently active	Occasionally active	Rarely or never active	Not authorized to deposit
Lecturers/Researchers					
Library staff					
Administrative Assistants					
Students					
Other (please specify)					

VI. REPOSITORY FUNCTION AND USE

15. Repositories are set up for a variety of reasons. Please rate the relevance of the following statements in relation to your repository objectives.

	Highly relevant	Relevant	Slightly relevant	Not relevant	I don't know
Enhance access to resources					
Promote new modes of publication					
Encourage new forms of peer review					
Aid institutional information management					
Promote data sharing					
Preservation of digital resources					

16. Electronic resources found within your repository are used:

[Tick one box only]

Frequently	Occasionally	Rarely	Never	I don't know

Depends on the type of material (please comment)

17. The use of the resources within the repository is monitored by: [Tick as many boxes as apply]

Myself
Library
Special committee
Computing services
Other
Not monitored
I don't know <please go to next section>

18. If the use of the resources in your repository is monitored please indicate how this is done [Tick as many boxes as apply]

Server transaction logs/ log analysis
Google analytics
Link analysis
User surveys/questionnaires
Download counts
Other (please specify)

19. Please indicate the extent to which you agree or disagree with the following statements. [Tick one box in each row]

	Strongly Agree	Tend to agree	Neither	Tend to disagree	Strongly disagree	I don't know
Repositories should only contain peer-reviewed resources						
Books and journal articles are the only valid form for communicating research						
A repository should contain a wide variety of electronic resources, such as images, datasets and software						
If university members are allowed to deposit any type of material in a repository, it will fill up with junk						
New types of electronic resources will change scholarly publishing						

20. Do you have any additional views or concerns about the types of electronic resources deposited in repositories and the different groups who deposit?
Your remarks would be greatly appreciated so please feel free to expand.

VII. ABOUT YOU AS REPOSITORY ADMINISTRATOR

Please enter information about yourself. These data will be anonymous in the results.

21. Name (First name, last name)

22. Your job position title

23. Would you be willing to be contacted by email if follow up questions are required? [Tick one box]

Yes

No

24. Would you be willing to be contacted for a short (20 min) interview to address similar issues in more depth? [Tick one box]

Yes

No

25. Please enter email

VIII. THANKS!

I greatly appreciate the time you took to fill out this survey. Please do not hesitate to contact me if you have any further comments or questions.

Isabel Galina Russell
i.russell@ucl.ac.uk

Please tick the box if you are interested in receiving the report on the results of this survey.

Annex 2: Email lists descriptions

This annex offers a description of the email lists to which the online survey invitation was sent.

OpenDOAR

An email distribution service that was set up as a bespoke email redistribution service to address repository administrators registered at OpenDOAR. The service is flexible as it can be configured to directly address a specific portion of the repositories (for example, repositories in a particular country). For the survey email all repository administrators on the list was selected. Emails are filtered by OpenDOAR staff.

Dspace

An email discussion list to ask questions or join discussions about non-technical aspects of building and running a DSpace service. Highly likely that a large number are repository managers.

UKCORR (UK Council of Research Repositories)

An email distribution list for UKCORR whose function is to serve as a professional, independent body to discuss and advise in relation to repositories.

JISC-Repositories

An email list which is part of the JISC Digital Repositories Programme designed to bring together people across diverse disciplines to coordinate efforts for the effective building on repositories, including research, learning, information services, institutional policy, management and administration, records management, etc).

JISC-CETIS Metadata and Digital Repository SIG

A national research and development service, for standards based e-learning whose aims are to advise, promote and represent at international standarization level.

CODATA (Committee on Data for Science and Technology of the International Council for Science)

An email list belonging to the committee which hopes to improve scientific data and management and use, it is mainly in German.

Sigmetrics

A virtual special interest group of the American Society for Information Science and Technology, Sigmetrics listserv discussion group covers bibliometrics, scientometrics and informetrics and metrics as related to the design and operation of information systems.

Annex 3: Sample invitation email for online survey

I am undertaking research into the use of electronic resources within repositories with particular focus on diverse content type as part of a PhD thesis at University College London. As part of my research I am currently looking for repository administrators (or anyone working directly with a repository in their institution) who would be willing to fill in an online survey about types and use of electronic resources and depositing behaviour. It should take about 15 minutes to complete.

The results of this work should provide us with further insight into the use of electronic resources within repositories and help to find appropriate methodologies to detect and evaluate their impact. It is vital to understand if and how these electronic resources are being used and to what extent are they important within the scholarly communication process.

The survey is available at: http://tinyurl.com/2b348a
A Spanish version is available at: http://tinyurl.com/2bromg

Further information about the project can be found at: http://www.ucl.ac.uk/~uczciga/repositories/index.html

Please feel free to forward this email to repository administrators that you think might be interested.

I greatly appreciate your participation. If you have any queries or comments please email me at i.russell@ucl.ac.uk.

Thank you, Isabel Galina Russell

Annex 4: Consent form for interviews

Use of electronic resources in institutional repositories **research interview consent form**

I the undersigned agree to be interviewed by Isabel Galina-Russell. I understand that the data from the interview will be used for the purposes of PhD research. Data derived from the PhD will be published in the final thesis and will appear in articles in refereed academic journals.

All comments will be anonymised as far as possible. In a situation where a comment is only comprehensible if associated with the name of a given repository, the interviewee will be contacted and asked permission for his/her name to be used and will be given an opportunity to check the accuracy of the quotation. Interviewees may also request that certain information should remain confidential and is not for publication.

In the light of this information I give my permission for data derived from this interview to be used in the research project.

Signed

Date

Print name

Annex 5: Interview guide- Repository manager

I. Background

In the survey you mentioned that the repository is X number of years old.

When was the repository first developed?

What is the history of the repository?

Who created it?

Who funded it?

II. Content intake workflows

In the survey you mentioned that you have X number of items.

How were the initial items acquired? (ie. How did they get into the repository?)

What mechanisms are currently used for incorporating resources?

What role(s) do repository staff play?

III. Depositing behaviour

Who are the most active depositors?

Have and how have these academics been motivated/encouraged to deposit?

Do you know what types of materials they deposit?

What are the main disincentives to depositing?

IV. Usage

a) In the survey you mentioned that resource usage is monitored using log analysis/ Google analytics / link analysis / user surveys / download counts / other.

Since when?

Who suggested this and why?

Why did you choose this method for monitoring?

Has it been useful?

What have you learnt?

Have you changed/altered things from looking at the usage data?

How have you publicized the repository?

b) In the survey you mentioned that resource usage is not monitored.

Do you have any type of experiences/evidence that shows the level of usage?

Are you planning to introduce any form of usage monitoring (such as log analysis/ Google analytics / link analysis / user surveys / download counts /)?

V. Future

What are the future plans for the repository?

Have you noticed any significant changes in academics attitudes?

Any additional remarks?