Institutional Repositories:

Investigating User Groups and Comparative Evaluation Using Link Analysis

Paul Wells

A dissertation submitted to the University of the West of England, Bristol in accordance with the requirements of the degree of MSc in Information & Library Management.

Bristol Institute of Technology, May 2009.

Abstract

The aim of this investigation was to look at user groups of institutional repositories. Past research on repository users has focused on authors and depositors at the expense of other users, and little is known about what types of user groups are associated with institutional repositories. This investigation used the research techniques of link analysis and content analysis to investigate links to institutional repository websites and determine what types of user groups are using repositories. These techniques were also examined for their use in providing a comparative evaluation of institutional repositories.

After an initial pilot study, four UK institutional repositories were selected for investigation. A link analysis was carried out using dedicated software. The results of the link analysis were then subjected to a content analysis to provide additional context.

The findings of the research were able to partially answer the research questions. Using link analysis alone it was not possible to gather detailed enough data to identify distinct user groups. When combined with content analysis, broad user groups were identifiable. The user groups shown in the results included those identified elsewhere in the literature, such as authors, academics and repository administrators. In addition, there was evidence of use by teaching and research related users, professional and public users. It was found that link analysis of institutional repositories was not suitable for comparative analysis, as results were more closely linked with the age of the repository than other factors. The results sample available for content analysis was found to be too small to produce suitable results for comparative evaluation, although a larger sample size would be able to overcome this in any further studies.

Acknowledgements

I would like to thank my supervisor, Debra Hiom, for her support throughout this project. I would also like to thank Michelle Jeffery for her continued support and patience, past, present and future.

I'd also like to thank McVities Hobnobs and Tea, for the part they have played in the completion of this investigation.

AUTHOR'S DECLARATION

I declare that the work in this dissertation was carried out in accordance with the Regulations of the University of the West of England, Bristol. The work is original except where indicated by special reference in the text and no part of the dissertation has been submitted for any other degree.

Any views expressed in the dissertation are those of the author and in no way represent those of the University.

The dissertation has not been presented to any other University for examination either in the United Kingdom or overseas.

SIGNED: ..... DATE: .....

Table of Contents

List of Figures and Tables

Word Count: 17,451

1. Introduction

This chapter introduces the topic of the investigation, sets out the research aims and questions, and provides a rationale for their study.

1.1 Rationale

Institutional repositories are a set of services offered by universities in order to capture the intellectual output of their academics (Crow 2002, Lynch 2003). Whilst other types of institution may offer repositories, the majority of UK repositories are linked to Higher Education institutions (Brody 2007). These are offered in the form of a web-based service, usually delivered through the university's library division or similar. They are an increasingly important tool in UK Higher Education for a number of reasons, including:

- Their use by Open Access advocates to increase access to scholarly publications,
- Their possible use by universities in support of research evaluation procedures such as the Research Assessment Exercise, and
- The hope that they will ease the Journals Crisis by reducing journal subscription costs.

The UK has the second highest number of institutional repositories in the world (University of Nottingham 2008a). As a service of growing importance to university library provision, the evaluation of institutional repositories is important to their continued improvement. There have been many attempts to evaluate different aspects of institutional repositories. As a continually developing set of technologies are used to fulfil this role, it is important to understand how repositories are being used. One aspect in particular that has been investigated is users of repositories. Most studies of repository users have focused on depositors of material, that is, authors and academics that create scholarly material. However, institutional repositories also have other user groups. In particular, material in the repository is collated and maintained by administrators, and accessed by students and researchers. These users are under-represented in the current literature.

Institutional repositories are currently exclusively a web-based technology, and as such have seen some application of evaluative investigation aimed at their integration and visibility in the World Wide Web, including the use of webometrics and link analysis. This investigation will draw upon such methodologies, as described in the literature review and research methods chapters, and direct them at a comparative evaluation of the web integration of institutional repositories. In addition, through the use of qualitative and

quantitative methods, the hyperlink aspect of the web's structure will be able to be revealed and categorised to indicate types of institutional repositories user groupings. This will build on and extend the use of categorisation in describing links according to types of websites and motivations for link creation that have already been used in link analysis methodologies. The extension of the usage of link analysis to infer user groups has not been described before in the literature relating to Library and Information Science studies, but similar methodologies have been employed in other fields, such as ethnography, which will influence this investigation.

Given the increasing importance of institutional repositories it is important to address the lack of information regarding repository users. Investigating institutional repository users will increase the information available to repository administrators, allowing better decisions regarding future development of repository services.

The research aims were developed iteratively as the project progressed. The initial idea was developed through a professional interest in institutional repositories, combined with key papers discovered at the beginning of the literature review, namely McKay (2007) and Zuccala et. al. (2006). In developing the research aims and questions it was recognised that there is a need for these to be precise and feasible (Ryan and Walsh 2006). This was achieved by limiting the scope of the investigation to UK institutional repositories, and including the proposed methodological approach in the aim and questions.

1.2 Research Aim:

This research aims to investigate the users of UK based institutional repositories through the use of webometric link analysis in identifying user groups and comparatively evaluating institutional repositories.

1.3 Research Questions:

- Can link analysis be used to identify user groups for UK institutional repository?
- Can link analysis be used as a comparative evaluation tool for UK institutional repositories?

2. Literature Review

2.1 Introduction

The subject area of the investigation which dictates the focus of the literature review is indicated by the research aims and questions. The aim of the research is to investigate the users of UK based institutional repositories through the use of webometric link analysis in identifying user groups and comparatively evaluating institutional repositories. The research questions that are derived from the research aims are:

- Can link analysis be used to identify institutional repository user groups?
- Can link analysis be used as a comparative evaluation tool for institutional repositories?

The literature review will attempt to address three aspects of the research subject area indicated by the research aims. These are:

- The origins of institutional repositories and how they are affecting their development,
- Relevant current research into institutional repositories, including webometrics,
- How existing research has impacted this investigation, and how this investigation will contribute to the professional literature.

In order to give the reader sufficient understanding of the area under investigation, there is an introduction to the origins, history and drivers of development of institutional repositories. This is intended to highlight discussions as to the purpose and audiences of institutional repositories. By looking at some of the common definitions of institutional repositories it is hoped that the focus of the research will be more clearly defined in an area that still contains uncertainties.

Current areas of research into institutional repositories relevant to the study will be highlighted, in order to set the investigation in the proper context, and show which aspects of institutional repository deployment require further investigation. In particular, investigation into institutional repository user groups through study and discussion will be compared to identify possible methodologies and critiqued to highlight investigative flaws. In addition, papers discussing or investigating evaluative methodologies for institutional repositories will be compared to give background to current evaluative practices and issues affecting development of new evaluation methods. The research area of webometrics will be introduced, before focusing on the methodologies involved in link analysis. Key papers and investigations into repositories, digital libraries and

academic related areas using link analysis will be contrasted to highlight areas needing attention in the research methodology.

The literature review will aim to place the current investigation within the existing research literature, both in terms of how the literature review has impacted the investigation, and how the research undertaken will contribute to the professional understanding of the subject.

2.2 Origins of Institutional Repositories

Jones (2006) traces the first development of the idea of a repository of scholarly publications to the early 1990's and articles discussing changes in scholarly communication from Gardner and Harnad. These were the first indications of unhappiness with traditional methods of academic publishing. The emerging technologies of File Transfer Protocol, gopher, and the World Wide Web were used to increase availability of scholarly material by lowering the barriers to distribution. The tradition of informal circulation of research articles in some disciplines was initially duplicated via the new technologies. This was followed by more formalised discipline-centred internet-based repositories of pre- and post-print articles, the first example being arXiv in 1991 (arXiv.org 2009).

Early discussion of the issues affecting scholarly communication involves mainly academics. Moves towards an institutionally focused repository come later, and is heavily influenced by librarians and their associates. The first published proposal for an institutionally focused repository was made by Okerson and O'Donnell (1995), writing for the Association of Research Libraries.

The development of stable open source software with which to implement institutional repositories is seen as pivotal to the rapid increase in their deployment (Jones 2006). The earliest examples of such programmes are Eprints, released in 2001 and DSpace, released in 2002. These built on the foundations set by electronic thesis software, such as ETD-db, available from 1999. One further development that encouraged the deployment of software was the development of the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH). This recognised a standard level of metadata required for digital repositories and enabled the automation of record-sharing between institutional repositories and secondary services (e.g. search engines, harvesters) to give Institutional repositories a wider audience (Ware 2004).

In a summary of the pre-cursive factors that led to the development of institutional

repositories, Jones lists the following elements:

- E-thesis archives
- Departmental e-print archives
- Faculty practice of e-prints on personal web pages
- Subject repositories
- Need from institutions for preservation/presentation of research output
- Open Access aims
- Distributed document servers
- The 'Journals Crisis'.

(Jones 2006)

This illustrates that due to their origins institutional repositories have many different factors driving their original and continued development. Groups associated with repositories, such as librarians, authors and 'archivangelists' (proponents of Open Access, Adams 2007) may have conflicting objectives. This suggests the ultimate success of institutional repositories will depend on the ability of users and managers to understand their differing objectives and synthesise solutions to satisfy their mutual aims. Jones notes the current lack of focus of institutional repositories:

> Institutional repositories tend to have a very wide remit. They mean many different things to many different people, and are used in a variety of ways (Jones 2006:114).

2.3 Definitions

Institutional repositories are firmly based within the theoretical framework of digital libraries. Jones et. al. (2006) sets out the inclusion of institutional repositories within digital libraries by first looking to define Digital Libraries, although ultimately finding no common consensus. Through comparison with Ranganathan's (1936) five laws of library science, Jones indicates that Digital Libraries can only be considered as within the traditional scope of libraries based on the condition of selection, i.e. materials are included in a collection subject to a collection development policy. This opinion is in common with institutional repository practitioners who see lack of clear collection policies as a barrier to further institutional repository development (Salo 2008). Jones et. al. (2006) also note the dilution of the phrase 'digital library' through common usage in computing. Heery and Anderson (2005) distinguish digital repositories from digital libraries in defining a digital repository as having the following characteristics:

- Content is deposited in a repository, whether by the content creator, owner or third party,
- The repository architecture manages content as well as metadata,
- The repository offers a minimum set of basic services e.g. put, get, search, access control,
- The repository must be sustainable and trusted, well-supported and well-managed.

There are several forms of digital repository apart from institutional, including learning object repositories and research data repositories (Zuccala 2007). Though all share common attributes, suitable definitions are needed to adequately distinguish between them for the purposes of function, administration and investigation.

There are several key definitions of institutional repositories that are widely quoted. In particular, Crow's (2002) definition is one of the earliest in the literature, and so is considered influential:

> Any collection of digital material hosted, owned or controlled, or disseminated by a college or university, irrespective of purpose or provenance (Crow 2002:16).

Similarly, Lynch's (2003) definition 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 the institution and its community members. It is most essentially an organizational commitment to the stewardship of these digital materials, including long term preservation where appropriate, as well as organizational and access or distribution (Lynch 2003:2).

The distinction between the two is important, being the emphasis on the collection of material in Crow's, and the services provided in support of access to materials in Lynch's. Both definitions together provide a reasonable summary of the essence of institutional repositories, as they are currently deployed in the UK; however there is still enough discussion surrounding institutional repository definitions for them not to fit in all cases. It is noteworthy that both Lynch and Crow wrote their works containing these definitions for the Association of Research Libraries, essentially marking the point at which librarians entered into discussions about repositories and scholarly communication. Even in these early documents it is possible to note the different approaches between the librarians on the one hand and the scholars on the other. The librarians' position is driven by the need to balance budgets in the face of the journals crisis, whilst the scholars are focused on adequate access to information.

Jones (2006) attempts to synthesise the key points of definitions of institutional repositories from the literature as being:

- Institutionally defined
- Scholarly
- Cumulative and perpetual (i.e. continuously open and available)
- Open and interoperable (Open Access and Open Archives Initiative compliant)
- Capturing and preserving events of campus life
- Searchable within constraints.

Despite the emphasis on universities within the definitions discussed, it is noted that not all institutions with related repositories are higher education establishments. However, the majority of institutional repositories in the UK are related to universities (Brody 2007), and this will inform the focus of this investigation.

Jones (2006) and Poynder (2006) suggest that institutional repositories are not yet well established, and uncertainties regarding definitions only serve to underline this. In differentiating between institutional repositories, digital libraries and other repositories, the only distinction given is the institutional focus of the service or collection. However, this does not preclude institutional repositories from serving a useful purpose and being seen as one tool amongst many in the scope of the digital library. Continued research and discussion addressing the underlying issues facing institutional repositories will be necessary if they are to fulfil the potential identified for them (Harnad 2001).

2.4 Technical Aspects of Institutional Repositories

With regards to technical aspects of institutional repository development, specifically the software and hardware used to run repositories, there is a surfeit of information (McKay 2007). There is a tradition in the literature of repository managers publishing case studies of their institutional repository deployment (Barwick 2007, Bevan 2007, Herb and Muller 2008, Jayakanth et al. 2008) including technical requirements of software, hardware and support. In addition, there are evaluative reports of key software programmes published by national and international organisations (Open Society Institute 2004, O'Connor 2006), and software user groups peer support through email lists and wikis (DSpace 2008, Eprints 2008).

The general consensus gleaned from the literature is that there are two key software platforms, DSpace, and Eprints, and several other lesser-used platforms. Discrimination between the two key platforms tends to be on grounds of preference or existing

technical abilities (University of Bath 2008). In addition there are organisations that provide a managed solution based on the open source software for a fee, for example Open Repository (Open Repository 2008), and Digital Commons (Berkley Electronic Press 2008) which can also offer additional related services, for example faculty liaison materials to encourage deposit. These managed services remove or minimise the need for in-house technical expertise when deploying a new institutional repository. This in effect means that the technical aspects of institutional repository development and deployment that dominated the experiences presented in the early literature have now been essentially sidelined, and creating a new repository is, from a technical point of view, reasonably straightforward.

This increased ease of implementation is reflected particularly in newer studies looking at institutional repository 'ecology' or "the interactions...between repositories and other systems, processes, and people" (Robertson et. al. 2008), which is reflected in the focus of this investigation. However, it is worth noting that institutional repositories are regarded by some as still in their infancy, both as a technology and a resource (Aschenbrenner et al. 2008), and so further changes of both the software and role of the repository is likely in the future, perhaps leading to further specialisation of repository functions and fracturing of the definitions of institutional repositories. Information regarding the evaluation of the performance of institutional repositories will therefore be necessary to ensure they are able to achieve the roles and targets set for them.

2.5 Open access

As outlined by Jones (2006) and others, one key factor behind support for institutional repositories is that of the Open Access (OA) movement for scholarly communication. Advocates of OA suggest the current scholarly communication process of publishing in toll-access journals is ultimately a barrier to efficient communication, and this can be overcome through author self-archive of electronic post-prints (e-prints) in author websites, institutional or subject repositories (Harnad 1999). This objective is similar to that of librarians responding to the 'journals crisis', the disproportionate rise in journal subscription costs compared with inflation (McGuigan and Russell 2008). Librarians advocate the use of self-archiving as a tool to combat the dominance of journal publishers (Crow 2002). However, as Poynder (2006) notes, OA advocates (authors) and institutional repository managers (librarians) may ultimately have different motivations, and highlights the distinction between affordability and impact. He quotes Harnad as suggesting OA advocates emphasise a complementary model that will co-exist with traditional scholarly publishing, where as advocates of institutional repositories wish to subvert or replace journal publishing.

As evidence of the complementary approach of OA, the Budapest Open Archive Initiative (BOAI) has produced guidelines on OA journal publishing and business models (BOAI 2004). However, Rowlands (2005) notes that in a survey of authors on attitudes to OA, both OA and Institutional repositories were thought to be undermining to traditional journal publishing. Even so, it would appear that the ultimate aim of both viewpoints is to increase the availability of scholarly material through reduction of barriers to access. OA is also important to institutional repositories as a promotional tool, by encouraging author deposits because of the OA citation advantage (Mark and Shearer 2006). However, this effect remains controversial (Davis 2008, Harnad 2004), and it is difficult to discern clear patterns (Xia 2008).

2.6 Institutional Repository costs

It has been noted that despite the development of Institutional repositories in response to rising journal subscription costs, no libraries have yet reported a reduction in costs as a benefit of installing an institutional repository (McKay 2007). Also, the costs of setting up and running a repository have traditionally fallen to the library service in addition to journal costs. Estimates of the actual costs of running a repository are difficult to produce due to the number of variables between institutions, but JISC estimates start up costs at £80,000 and annual running costs (including staffing) of £40,000 (JISC 2005). In contrast, Houghton et. al. (2009) estimated that in evaluating the costs and benefits of alternative scholarly publishing methods (toll access publishing, open access publishing and author self archiving) there was still a considerable cost saving in author self archiving (i.e. institutional repositories) versus OA and traditional publishing, equivalent to approximately £1,180 per article. This is, however, a long-term view, and does not currently reflect actual library costs. The amounts of money involved in implementing an institutional repository are therefore quite substantial, and will need justification if it is to continue. It remains to be seen if funding for these costs is sustainable. Salo (2008a) has suggested that currently this is not the case, mainly due to well-observed factors, such as low deposit rates and lack of faculty interest (Davis and Connolly 2007). In addition Salo (2008b) highlights that these issues are causing disillusionment to librarians and administrators of institutional repositories.

2.7 Repository Users

The importance of studying institutional repository users is highlighted by Schmitz (2008), who notes that "understanding use [is] a pathway to sustainability". Through research into how institutional repositories are used and who uses them, they can be developed to provide a better utilised and more responsive service.

There is an emphasis on authors as depositors in the research literature regarding institutional repository users (Schmitz 2008, McKay 2007). This probably reflects that when they are first deployed institutional repositories need to recruit relevant content to provide their services (Thomas and MacDonald 2007). However, depositors are not the only users of institutional repositories, and academics use repositories as both depositors and researchers. Rowlands (2004 and 2005) offers some insight into authors' attitudes to institutional repositories in their dual role as depositors and users. Authors give conflicting views on certain aspects of scholarly publishing, particularly in relation to their satisfaction with their access to journals and their dissatisfaction with the cost and proliferation of journals. More authors report using their own website to host their output than those using a repository. Differences seem to be mainly a product of the age of the author, with younger academics being more likely to be aware of and have positive views of OA and repositories. In addition, whilst the application of these studies to current attitudes may not be relevant, it is positive to note that between the two studies (2004 and 2005) the awareness of OA issues was measured to have risen.

The most insightful discussion of repository users beyond authors is by McKay (2007), who addresses the usability of institutional repositories by three distinct user groups; Authors, Information Seekers and Data Maintainers. Of the three, McKay suggests Information Seekers (or end users/researchers) are most neglected in the literature, and if institutional repositories are to fully realise their potential then this must be addressed. McKay attempts to gain insight into this group of users by comparing research approaches into information seeking in similar contexts. The comparative methodology used by McKay could be useful in discovering institutional repository user groups by highlighting which groups use similar resources, such as OA journals. However, as McKay notes, a more direct method would be preferable, to give directly relevant data with which to draw conclusions. In addition, McKay highlights the importance of search engines to institutional repositories; end users must be aware of the institutional repository's services in order to make use of them, and the most effective way of increasing visibility is via search engine indexing. The comparative approach of McKay is also used by Schmitz (2008), again in response to the absence of sufficient useful studies. Schmitz looks at digitization projects and institutional repositories, and notes that prospective user groups are often ill-defined, including students, scholars, the public and worldwide users. Both McKay and Schmitz highlight the importance of knowing the status of repository in order to evaluate other aspects of the success of the repository, and that there may be users of repositories that administrators are currently unaware of.

No investigative studies of institutional repository users and no studies of repositories involving users outside of the institution were found in the literature. Following the

example of McKay and Schmitz above it is reasonable to look at investigations into similar services for information regarding user groups.

As noted in the discussion of definitions, institutional repositories can be considered as a sub-set of digital libraries, and are generally closely linked to other library services, and so literature relevant to these areas was also examined.

In discussing users associated with academic libraries, Brophy (2005) identifies 16 stakeholder groups, although of these only those related to the roles of students, academic staff and the public are likely to be users of library holdings. However, investigation into user groups is lacking in the literature, perhaps as users of these services are considered to be self-evident.

Similar investigations into digital libraries also appear to be limited. Fuhr et. al. (2001) identify the user categories internal, general, education, professional and research, when developing an evaluation criteria for digital libraries. No further relevant work was found relating to other specific types of digital repository during the literature review. This may be due to a concentration on user demographics as opposed to user groups (Cherry and Duff 2002).

2.8 Evaluating Institutional Repositories

Performance indicators are used to evaluate how well an organisation or project is meeting its expected targets. Ideally, standard performance indicators will be used across related organisations. However, in new areas of practice there may not be enough evidence to validate the use of a particular set of indicators or tools. This is certainly the case with institutional repositories (Kim and Kim 2006). This lack of common methodology for evaluation is reflected in institutional repository literature aimed at suggesting and evaluating methods of evaluation for Institutional repositories. In particular, Westell (2006) proposes a series of qualitative measures designed to evaluate different areas of institutional repository implementation that have been based on Canadian institutional repositories. Fuhr et. al. (2007) note three kinds of evaluation; formative, carried out in parallel with development, summative, carried out after an initial release, and comparative, whereby systems and components are evaluated against each other. Although not implicit in many investigations, the literature relating to established institutional repositories is largely comparative.

Thomas and MacDonald (2007) summarise a number of both qualitative and quantitative measures proposed in the literature, before outlining a framework of performance

indicators for different institutional repository functions (i.e., inputs, outputs and impact). In particular, criticism is levelled at a 'bean-counter' outlook on evaluation, where quantitative evaluation methods are used without critique. Particularly, attention is drawn to Carr and Brody's (2007) investigation of a 'sustainable deposit' profile, confirming that in assessing performance indicators more attention is paid to authors/depositors than information seekers. In Thomas and MacDonald (2008), they go on to discuss the possible future evaluative measurements of institutional repositories, suggesting that usage and impact will be important evaluative factors. However, no suitable tools to achieve such an evaluation are presented.

Zuccala et. al. (2006, 2007, and 2008) reports on an earlier project aimed at evaluating repositories with a mixed methods approach. The study aimed to examine management of a range of digital repositories, through interview with managers, a questionnaire survey of users and a web-based link analysis study to illuminate user groups. The discussion with managers highlights some of the methods already in use to identify user groups, but focus on depositor users of repositories, in common with other studies. The user survey is an interesting insight into repository users but again focuses on already visible user groups, that is, those who could be identified and contacted for the survey. The most useful but hardest to interpret is the web link analysis. The use of web links to highlight hidden user groups is one of obvious benefit, especially for comparative evaluation of different institutional repositories where accessible institution specific information may not be available. Link analysis has also been proposed as a comparative evaluation tool for website managers (Thelwall 2009a).

The use of web-based evaluation tools is appropriate when regarding the fact that institutional repositories are digital collections, and therefore inherently online resources (Crow 2002, Lynch 2003). In particular, as part of the web they are tightly linked with search engine technologies. Hitchcock (2003) suggests that "the search engine has become the de facto interface to information", a quote supported continually by user surveys and information seeking behaviour research. In relation to repositories, research papers such as Markland (2006), which looks at how available institutional repository articles are via Google, and case studies such as Organ (2006), that states Google as being identified as the primary access and referral point for an institutional repository, re-enforce the importance of search engines to repositories, and emphasise that institutional repositories are a web-embedded technology. So we can see that search engines are important points of discovery for institutional repositories, and an understanding of how search engines direct users to repositories is useful. In contrast, the email survey component of Zuccala et. al. (2006) reports the majority of respondents claimed to discover the institutional repository via colleagues, and a

negligible number via search engines. This may point to a lack of institutional repository impact when users are being referred from a search engine.

Institutional repositories could be considered as part of the 'deep web', that part which is difficult to find due to its inaccessibility to search engine indexers (Bergman 2000). Search engines are of increasing importance in research discovery, particularly for institutional repositories (Markland 2006). Increased efforts are therefore being made to improve the visibility of institutional repositories to search engines, particularly through use of Open Access Initiative Protocol Metadata Harvesting, though McCown et. al. (2006) show this has had varying success.

In summary, evaluation of institutional repositories is currently not standardised, and is generally comparative. It is recognised that repositories are exclusively accessed online, and so an examination of evaluation using internet relevant methodology is appropriate.

2.9 Barriers to Future Success

Jones (2006) notes some of the key points regarding development of institutional repositories. Notably, that they are "old enough in concept, [but] still young in implementation" and

> If the institutional repository does not yet inhabit a defined place in the information environment, then they are not sufficiently well established to even be considered essential elements (Jones 2006:116).

Aschenbrenner et. al. (2008) discuss overall institutional repository adoption in terms of expectations, firstly being over inflated by promise, then troughed by disillusionment, before reaching a plateau of productivity. However, this enlightened ending is far from guaranteed.

The difficulties which are affecting the successful establishment of institutional repositories are the same ones outlined in the earliest discussions, namely, how to replicate the peer-review process (quality control) and the perception of print publishing as having authority that electronic publishing does not (Okerson and O'Donnell 1995). In addition, Wilson (2008) notes that whilst the details of the publishing process have been affected by technology, publishing and usage models in scholarly communication are still derivative of the print era. This implies that the future for scholarly publishing in general is uncertain.

One key aspect to the future success of institutional repositories will be their ability to fulfil the promise and potential described in the definitions from Crow, Lynch and Jones

discussed above. For example, Jones' "cumulative and perpetual" (2006) requirement is a criterion that will only be tested over time, and doubts currently exist over the archival potential of all digital libraries (Seadle 2008). In particular there are already doubts over the ability of institutional repositories to fulfil this function in their current state (for example, see Hockx-Yu 2006).

This discussion serves to illustrate that the future of institutional repositories is by no means secure, and suitable comparative evaluation coupled with an understanding of who are using the repositories and why, will be needed to ensure their long-term viability.

2.10 Webometrics and Link Analysis

This section should serve as an introduction to the area of research known as Webometrics, and the techniques of webometric research known as Link Analysis. A discussion of the merits of link analysis in pursuing the research questions, and possible alternative methodologies, is presented in the research methods chapter.

Webometrics is:

> The study of the quantitative aspects of the construction and use of information resources, structures and technologies on the Web drawing on bibliometric and informetric approaches (Bjorneborn 2004:12).

Originally defined by Almind and Ingwersen (1997), the field arose from the application of bibliometric analysis tools, used in relation to journal article citation, to new forms of electronic communication, particularly in relation to scholarly communication. Webometrics encompasses techniques and research from a range of fields. Both bibliometrics and webometrics are considered to be sub-divisions of informetrics, or the study of quantitative aspects of information (Bar-Ilan 2008). The term webometrics can be applied more widely, to encompass all quantitative analysis of web-related information, analogous to web dynamics in computer science (Thelwall 2005b).

The two most widely used webometric analysis tools are link analysis and web log file analysis (Thelwall 2007a). There are multiple ways in which these techniques can be applied to research, which can make designing the appropriate approach difficult. A further difficulty is a lack of well-defined vocabulary, which often affects new, cross-disciplinary or loosely defined fields of study (Kennan and Wilson 2006). This is illustrated particularly by the use of the parallel term cybermetrics. According to Bjorneborn and Ingwersen (2004) the two are distinguished by their research focus.

Webometrics is concerned with the study of quantitative aspects of the World Wide Web, whilst cybermetrics concerns study of all internet technologies, including email, file transfer etc. As webometrics is rooted in bibliometrics and informetrics there is considerable overlap with similar research from alternative backgrounds, particularly computer science (Thelwall et. al. 2005). Thus it is possible that alternative methodologies exist to address the same research issues from different research perspectives.

There are important differences between bibliometrics and webometrics, particularly in the structure of their environments. Bibliometrics studies the highly rigid structure of citations between scholarly journal articles, whilst webometrics studies the much more fluid, informal and contextual hypertext links between web pages. This lack of fixed structure between web pages has caused some authors to doubt the validity of webometric research in general (Thelwall 2007a), but much research has been carried out to assess the validity of conclusions drawn through webometric research.

Payne and Thelwall (2007) investigated the stability of website size and the number of links between them over time, drawing the conclusion that these properties stabilised over time, implying that webometric studies may have long term validity. However, there are alternative explanations for this result, including the increase in dynamic web pages excluded by webometric methods, and websites that are obsolete but not removed. Kousha (2005) reviewed investigations into correlations between results of webometric techniques and other methodologies, including university rankings and research productivity measures, concluding that there are strong correlations between such measures. Vaughan and Hysen (2002) were amongst the first to be able to show correlation between web impact factors and traditional journal impact factors. This persistence of web-structure, linked with correlation to external measurements, strongly implies that results from webometric investigations can be used to draw reliable conclusions. However, studies from an earlier period of the development of the web failed to find evidence of such correlation (Thelwall 2001, Harter and Ford 2000), indicating that such correlations may be related to the structure of the web as it exists currently, and that should this change, the correlation may not last.

Webometrics and link analysis are usually used exclusively on web pages. However, they have also been used to identify alternative types of documents that are available on the web (Thelwall and Kousha 2008). This was found to be important in the investigation, as non-web page documents were found to be have links to institutional repositories.

2.11 Link Analysis, Institutional Repositories and Users

Webometric research has a strong tradition of investigating academic trends and library-related phenomena (Thelwall 2004). Institutional repositories have been subject to a small number of webometric link analysis studies (Zuccala 2006, 2007, 2008). In addition, an annual ranking of digital repositories worldwide is undertaken by the Cybermetrics Lab using webometric techniques (CSIC 2009). The small number of studies of institutional repositories with link analysis methods is probably in part due to the timescale; institutional repositories, particularly in the UK, have only become established in the last 4-5 years, whilst webometric research, particularly link analysis, requires well established websites to give suitable results (Thelwall 2009a).

Within webometric link analysis there is a focus on investigating motivations for link creation, and correlating linking with related factors. This is perhaps due to webometrics origins in bibliometrics and citation analysis. Existing studies have looked at classifying the types of sites that links originate from, but have not extrapolated the types of users who are linking to a resource. For example, Belden (2006) uses link analysis to investigate links to and from the websites of special collection libraries, categorising results by website type. There is no discussion, however, of user or user groups who might be identified as producers or followers of links, as link analysis at that time focused on measures of web-visibility. Other studies have looked at motivations for creating links, but not which users are being motivated to create links. For example, Wilkinson et. al. (2003) looks at motivation underlying the creation of links between academic websites, but does not discuss who is creating these links, and whether they represent a specific type of user. This is perhaps less important when it is assumed that all users will be within the target group studied (in this example academics), but there may be many groups using institutional repositories that administrators are currently unaware of.

Zuccala (2006) presents the methodology of link analysis as a suitable tool for the evaluation of digital repositories. As is discussed in the text, and elaborated in the 'web-intelligence reports' (Zuccala and Thelwall 2005) the evaluation comprises two main parts:

- Identifying link-motivation, i.e. the reason for an individual or organisation to endorse a particular webpage,
- Identifying possible user groups revealed by links to web pages that may have been overlooked in assumptions about users.

However, the move from specific website examples to general assumptions about user groups is not made in this research. In addition, the report uses co-link analysis and statistical analysis to map link relationships graphically, an approach that is not easy to replicate.

Some link analysis studies appear to make several assumptions regarding relevant links. In particular that:

- The only relevant links are academic related, or
- The perceived motivations for creating a link are more important than information regarding who made the link, or
- The relationships reflected by the link are more important than the users creating the link.

(Thelwall 2003, Wilkinson et. al. 2003)

Other researchers using link analysis have assumed specific links do represent a user, who may be representative of a particular user group. Schmitz (2008) notes the possible use of field experiments and online ethnography as possible tools for investigating institutional repository users. Beaulieu (2005) uses link analysis methods to identify hyperlinks, and ethnographic methods to investigate these as users. Having established external pages linked to an online resource, several aspects were examined to determine user groupings. These were related to the context of the link within the web page, and included the 'nature' of the website, the intended audience (the user group) and the visual context of the link (e.g. positive or negative presentation etc.). However, the shortcomings of this approach are highlighted, including the limitations of gathering links via search engines and the nature of the web itself, but also in determining categories for such things as a websites 'nature', and what constitutes a user group. Schmitz (2008) also notes the difficulties in gathering large amounts of qualitative data, and recommends linking these methods with automatically generated data, such as webometric results.

The use of content analysis as a complementary tool to link analysis is well established in the relevant literature (Thelwall 2009a). In particular, a number of studies have sought to show validity of content analysis categorisations through the use of multiple classifiers (Vaughan et. al. 2006, 2007), though this has not always been successful (Harries et. al. 2004). Validity in content analysis is important to show the ease of replication of results (Weber 1990). The use of inductive categorisation has been shown to be effective in similar investigations (Vaughan et. al. op. cit.).

As discussed earlier there is currently very little research into describing repository users, and while certain assumptions exist as to who they are, there is still value in exploratory research providing an initial illumination, even if the conclusions must be placed within the limitations imposed by the research methodology.

3. Research Design

3.1 Introduction

The purpose of the research design chapter is to describe and discuss the methodology used to address the research aims and questions of this investigation. The aim of this project is to investigate the use of webometric link analysis in identifying user groups and evaluating institutional repositories. It will do so by answering the research questions:

- Can link analysis be used to identify repository user groups?
- Can link analysis be used as a comparative evaluation tool for institutional repositories?

It will address how the choice of subject and methodology was arrived at, how a webometric link analysis approach was undertaken through the use of automated link analysis software, and why this was thought to be the best approach. It outlines how the use of a pilot study influenced the research methodology, including the sampling method and content analysis process. It describes how suitable institutional repositories were selected for the link analysis process. The final part of the chapter addresses possible ethical issues raised by the methodology, as well as issues of validity and reliability of results.

3.2 Literature Review

The origins of the investigation arose from a personal interest in institutional repositories, fuelled through my employer deploying a new institutional repository. The initial literature review, the first step in synthesising the research approach, revealed the main themes that were to influence development of the investigation. A thorough review of appropriate literature, with a strong emphasis on professional and research papers, was undertaken into the areas surrounding the research questions. A large volume of work has been published on institutional repository research, and appropriately, much of it is available via OA journals, subject and institutional repositories.

The main themes highlighted in the literature review are:

- Institutional repositories are at an early stage of their development, as evidenced by lack of clarity of definitions and evaluation tools.
- Existing research has focused on technical issues, with less focus on institutional repository users.

- Of the research undertaken into users, most focuses on depositors and maintainers of material, with very little research into information seekers.

3.3 General Approach

A webometric link analysis research approach was decided on being the most appropriate method. This methodological approach has been used in a number of previous studies, particularly in relation to scholarly communication, and is discussed in more depth in the literature review. The main advantage of a link analysis approach is that it is possible to gather useful information on institutional repositories through the use of web search engines without requiring access locally administered IT systems. Had another webometric approach or analogous methodology been used it may have been necessary to have direct access to institutional repository IT systems. An exploration of alternative methodologies that were considered is given in the section below.

Link impact reports were created using a link analysis approach in order to illustrate the types of websites linking to institutional repositories. Automated software was used instead of completing a link analysis study manually, as this would have taken considerably more time. The software used was LexiURL searcher, which is a development of and replacement for LexiURL (LexiURL Searcher 2008). LexiURL has previously been used to investigate online scholarly communication in informal settings (Wilkinson et. al. 2003) and has been presented as a tool for investigating use and users in digital libraries (Zuccala et. al. 2007).

Webometrics in general and link analysis in particular have been criticised for relying on uncertain methodologies. For example, assumptions over the reason for links, the shifting nature of links between web pages, the inability to know how search engines calculate the relevance of links etc. It should be acknowledged that all research into web-based phenomena would be subject to these uncertainties; however this doesn't necessarily invalidate the methodology. By understanding the difficulties and making them explicit, valid conclusions can still be drawn from the investigation. In addition, through the evaluation of webometric techniques against more verifiable methods, an idea of how reliable the results can be is given. This is discussed further in the literature review chapter.

3.4 Alternative Methodologies

There are several methodologies that have been employed in previous studies either complementary or alternatively to webometrics and link analysis that could give insight

into the research questions. In particular, web-log analysis, questionnaires and interviews, and case studies will be discussed. In addition, some other software packages are available for link-analysis, and justification for using LexiURL searcher will be given.

Web-log analysis is a tool which has been used successfully in a number of institutional repository and related studies. In particular, CIBER (2008) used the tool to investigate usage of a wide variety of users' information seeking behaviour in digital environments. Nicholas et. al. (2006) used 'deep' web log analysis to investigate users and usage of digital libraries. Thelwall (2009) recommends web log analysis as complementary to link-analysis investigations, but also draws attention to the ultimate reason this methodology is unsuitable for this investigation; access is needed to the log data files, which is only available to webmasters. In addition, information regarding user attributes is limited in log files. Carr et. al. (2008), in discussing institutional repository statistics, note that web log analysis can be considered unreliable when used inexpertly, and suggest that a more nuanced approach is needed to determine the 'academic usage' of papers in a repository. This could in part be provided by a link analysis approach.

Questionnaires and interviews are similar common techniques used in library and information science investigations. Pickton (2005) utilised both techniques in discovering managers' and research students' attitudes towards Institutional repositories. Zuccala et. al. (2006) use both an email questionnaire and face-to-face interviews as a companion to link analysis in their investigation of digital repositories. Thelwall (2009) again recommends the use of these techniques to complement link analysis results. Creswell (2003) identifies the drawbacks of interviews as biases introduced by both the interviewer and interviewee, whilst Rugg (2007) notes several difficulties in adequately deploying questionnaires, suggesting that they should mainly be used as a supplementary method. Nicholas et. al. (2007) also note difficulties in differences in understanding technical terminology, which complicates this type of method. In addition, this methodology would struggle in answering the research questions, in particular due to the lack of existing knowledge of institutional repository users identified in McKay (2007), that is, it would be uncertain who to ask questions of; this difficulty is encountered in the research approach of Zuccala et. al. (2006).

Case studies are commonly used in investigations of institutional repositories. Examples involving institutional repositories include Bevan (2007) and Barwick (2007), describing specific institutional repository projects. A broad definition of case studies would include many link analysis studies, the most relevant being Zuccala et. al. (2007). Typically, a case study involves using multiple methods on a single instance of the phenomenon

under investigation. This was thought to be at odds with the need to generate comparative data to validate assumptions about users across a sample of institutional repositories.

Whilst all the additional methodologies mentioned above could have contributed to this investigation, it was ultimately decided that link analysis combined with content analysis would be most suitable to answering the research questions. The specific strength of the combined approaches is that no direct access is required, so the research can be carried out at a distance to the repositories under investigation. This is elaborated on in the sampling and access section. In addition, the combination of qualitative and quantitative approaches has been discussed below as beneficial in the validity and reliability section.

In addressing alternative link analysis software it is important to note that as the internet and World Wide Web grow in significance both for commerce and research, the number of tools dedicated to its analysis grows. These can be broadly split into those of commercial focus and those of academic (research) focus. It was desirable in this study to use a tool developed for academic research purposes. It is difficult to identify relevant analogous programs discussed in the literature, in part because of a lack of standardised terminology. Thelwall (2009) highlights three alternative link analysis software programs, LexiURL searcher (LexiURL 2008), Virtual Observatory for the Study of Online Networks (VOSON 2008) and issue crawler (Govcom.org 200?). Of these, LexiURL searcher is the obvious choice for this study, as the others are predominantly used for crawler-based surveys, requiring more consideration of ethical implications, discussed below. In addition, when conducting the literature review more information on the usage and previous applications of LexiURL was available in peer reviewed output, particularly with reference to digital libraries, suggesting its usage is more widespread in the area of library and information science. Finally, it appears to the author that the different programs available are subtly influenced by their originating communities. For example, all three programs mentioned here are described as social science tools, but LexiURL searcher is particularly identified as linked to Library and Information Science research (Thelwall 2009a).

3.5 Pilot Study

In order to confirm that the identified approach would give meaningful results, a short pilot study was undertaken before the main research commenced. A link impact report was created for the first repository identified for investigation using the LexiURL searcher software. A content analysis of the reported links was then undertaken to further analyse aspects of the data gathered relating to the research questions.

Piloting the research methods served to establish that reasonable results would be gained, and that the method was achievable. Prior to undertaking the pilot study it was uncertain whether results could be gathered using LexiURL searcher, and how extensive these results would be. The literature review highlighted a number of papers that described similar investigations, but it was important to validate this before committing to the full research project. The institution chosen for the pilot study was done so using the criteria discussed below. In addition, as the institution had been previously included in a webometric link analysis study of digital libraries by Zuccala et. al. (2006) it was expected that sufficient data could be gathered successfully. A description of the profile of all the institutions investigated is given in the findings chapter.

The pilot study also allowed the researcher to gain familiarity with and understanding of the software used, and the processes and results involved in a link analysis study, as outlined below. In addition, it was useful in estimating the time required for gathering the data needed for the investigation, and hence the number of repositories that could be investigated in the time available. In particular, in contrast to the web intelligence reports produced by Zuccala et. al. (2006, 2007, and 2008), it was found that it was not possible to investigate and analyse site co-links. This was due to the need to use additional software unavailable to this researcher, and the additional time such analysis would have taken.

The results of the pilot study were also used as a starting point for the content analysis. In combination with the literature review the results were used to inductively determine the categories for the content analysis undertaken and outlined below.

Certain aspects of the results of the pilot study were not expected. The results highlighted that the web pages that were retrieved would include links from foreign language websites. These would not be suitable for certain aspects of the content analysis discussed below as no interpretation of the content could be given. They were therefore not included in the pilot study. However, later in the investigation it was decided to include details of foreign language websites in recording the types of web pages recovered, in order to illustrate recognition of their importance in search engine results. In addition, some of the results retrieved were not web pages but other types of documents containing hypertext links to the repository. These documents were included in all analyses in line with previous studies identified in the literature (Thelwall and Kousha 2008).

The results of the pilot study are attached in the appendices.

3.6 Data Collection and Analysis

In order to fulfil the research aims, two types of analysis were needed. As discussed below, link analysis can highlight links to the identified website, but some additional analysis must be carried out to establish the possible user groups and link intentions. Thelwall (2009) suggests that a formalisation of random sampling of results to evaluate linkage motivations, using an inductive content analysis approach, is the best strategy. However, it also notes that content analysis in link analysis studies is generally undertaken to provide context to the results, rather than to accurately distinguish between categories. This is in part due to the amount of time and expertise an in-depth content analysis would require.

3.6.1 Link Analysis

The raw data was collected from search engines using the LexiURL searcher software, provided freely to researchers by the Statistical Cybermetrics Research Group at the University of Wolverhampton (LexiURL searcher 2008). Gathering data from search engines has implications for validity and reliability, discussed below.

LexiURL generates a link report by submitting to the selected search engine the search query:

linkdomain:www.site.com-site:www.site.com

where 'www.site.com' is the web address of the institutional repository web site under investigation. The search engine interprets the query by searching for all in-links (links directed at a page) to the web site, but removing all internal links from the same site. The link impact report generated by LexiURL searcher from the results returned by the search engine contains several parts. These comprised:

- Overall summary of results
- Complete list of matching web page URLs
- Matching web pages summarised by:
    - Domain
    - Site
    - Top Level Domain (TLD)
    - Second Top Level Domain (STLD)
- Random sample of web pages from unique domains (i.e. from different websites).

The aspects of the link report most useful to this investigation were the overall results

summary, results summarised by TLD and STLD, and the random sample of web pages. These were used compare the numbers and types of domains with links to institutional repositories, and the random sample was used as the basis for the content analysis.

The random sample of web pages is generated automatically by LexiURL searcher in two steps. Firstly, a random number generator is used to select up to 100 domains from the summary of web pages by domain. Then, for each domain name another random number generator is used to select a single web page as representative of that domain (Thelwall 2009b).

3.6.2 Content Analysis

Content analysis is used to classify a text according to words sharing similar connotations (Weber 1990). There are many methodological variations in applying content analysis (Weber op. cit.). In this investigation content analysis is used to uncover contextual information regarding links to institutional repositories from the web pages containing the links. This is in common with several other link analysis studies (Vaughan et. al 2007, Wilkinson et. al. 2003).

In order to answer the research questions, four different aspects of the links were analysed. These were:

- Source Pages
- Target Pages
- Link Motivations
- User Groups.

As there was no prior research examining links to institutional repositories, as part of the content analysis it was necessary to develop categories in order to classify characteristics of the web pages analysed. These were derived from a synthesis of previous studies identified in the literature as relevant, and an iterative process based on the data gathered. Similar approaches are described in Vaughan et. al. (2006) and Orme (2007). In particular, preliminary categories with brief definitions were created based on the sample of links examined, which were then added to and their definitions refined until the existing categories described all the links.

At first it was uncertain whether to include other types of documents in the content analysis. However, their inclusion in the link analysis results showed they contribute towards search engine ranking (Brin and Page 1998), and it was noted that other link analysis investigations have included non-web pages in their investigations (Thelwall and

Kousha 2008).

3.7 Sampling and Access

The method employed in gathering a suitable sample of institutional repositories was developed adaptively based on several previous repository studies. In identifying a suitable set of institutional repositories a number of decisions were made to limit the eligibility of sites. Particularly, this study was limited to the evaluation of UK based institutional repositories, based at the institutional level (i.e. not governmental, aggregating or disciplinary), with a multidisciplinary deposit profile, with deposits in English and deposits of articles (hopefully of a scholarly nature). These criteria are based in part on the focus of the study and partly on the assumptions present in the definitions discussed in the literature review chapter (i.e. UK institutionally focused repositories of a scholarly nature). In making this initial selection, use was made of the Directory of Open Access Repositories (OpenDOAR) (University of Nottingham 2008a), which lists repositories by such criteria. This limited the selection of institutional repositories to 65. In addition, pilot repositories or those set up too recently to have established themselves on the web were eliminated, and the remaining candidate repositories were ordered by size (that is, number of records contained). In addition to the author's criteria, the OpenDOAR website also eliminates candidate websites including those that contain no OA material or only references to documents, and sites that require log-in or subscription to access (University of Nottingham 2008b). The two criteria that were assumed to have the greatest impact on web visibility were size (number of deposits) and age (date established). To establish the age of institutional repositories two services were used; the Registry of Open Access Repositories (ROAR) (Brody 2007) and the Repository Records Statistics (Keene 2008). These allow ranking of repositories by age, which were then cross-referenced with the list of repositories ranked by size to give a combined list. In common with other evaluative investigations of institutional repositories it was hoped to include examples running on different software platforms (Kim 2006). Other institutional repository listing websites are available but do not have the search functionality needed for this investigation. In addition the institutional repositories identified for study were compared with the CSIC (2009) world-wide digital repository ranking. This established that the institutional repositories would have sufficient web presence to try and ensure good results. The combined list of anonymised repositories included in the investigation is described in the Findings and Results chapter. The number of repositories investigated was decided with reference to the relevant literature (Thelwall 2009a) and available resources, most notably the time available for the investigation.

3.8 Ethical Issues

Ethical issues can have an impact on webometric analysis investigations. In the case of link analysis through search engine interrogation, the information retrieved from the search engines is already publicly available. The search engine used, Yahoo!, limits the number of automatic searches that can be run in a 24 hour period. This avoids any strain on the search engine's operation through overuse by automatically generated searches, such as those used by LexiURL searcher. This limits the impact of the study on the search engine resource. In the qualitative evaluation of individual links from web pages, these again can be considered to be in the public domain, freely accessible via the web. For these reasons, special ethical considerations in this investigation were not considered necessary.

As the institutional repositories are not directly queried, it was not thought necessary to contact them prior to gathering the research data. As permission to use the repositories was not sought it was decided to anonymise the results to avoid any possible ethical implications in the discussion of these resources. A description of each repository investigated is included in the Findings and Discussion chapter.

3.9 Validity and Reliability

There are two issues affecting the validity and reliability of all link analysis investigations relying on search engine interrogation. These are limits on completeness of search engine results, and lack of transparency of search engine algorithms.

There are three major search engines that can be used by LexiURL searcher, and each is different in the way it responds to queries. The number of results returned to LexiURL searcher from the search engine is limited, and the number of results reported is likely to be a fraction of those indexed by the search engine in total. It is estimated that in a typical automated search, only around 10% of the total links will be found by LexiURL searcher. In addition, the LexiURL searcher program uses the search engines' Automated Program Interface (API), which can return dissimilar results to using the search engine via its web interface.

Although the basic tenets of search engines are known (Brin and Page 1998), the exact workings of search engines, how they find and rank pages, are commercial secrets. Much webometric research goes in to investigating the workings of search engines (Thelwall 2008a, 2008b). This introduces uncertainty into the use of search engines for academically rigorous research. However, as Thelwall (2008b) notes, "commercial search

engines are the only choice for some [webometric] applications".

The actual choice of search engine used is further limited by its availability and the search queries it supports. The three search engines supported by LexiURL searcher are Google, Yahoo! and Microsoft Live Search. Google requires a developers API code to gain access to its automated search features, and no longer makes these generally available, whilst Microsoft Live Search has withdrawn support for many of the automated search queries (Thelwall 2007b). This means that the only major search engine that gives readily available comprehensive results to automated search queries is Yahoo!. However, this does not invalidate the research method. It is also worth noting that Yahoo! was found to have the best coverage of OAI-PMH archives (Institutional repositories and DLs) of the three main search engines, although coverage of individual sites varies (McCown et. al. 2006).

Triangulation is a common way to confirm the validity of a study in the social sciences (Blaxter 2006). It usually takes the form of employing multiple methods in the research process, or "methods triangulation" (Patton 2002). This is also known as Mixed Methods Research, and is identified as being an increasingly important trend in Library and Information Science research (Fidel 2008). There are several possible ways that this could have been achieved in this study. Following the recommendations of Thelwall (2009), the quantitative approach of webometrics is complemented by the qualitative approach of content analysis. It is also important to recognise that triangulation cannot compensate for flawed methodology, and it is important to understand what is being studied, especially in online research (Jankowski and van Selm 2005).

In this study, content analysis will be employed in the analysis of the results to attempt to gain insight into the reasons for making the link (i.e. the link context), and to highlight the type of user that is linking to the institutional repository. Content analysis can take a number of forms, and so the style of the analysis will take its cue from Thelwall's (2004, 2009) recommendations given alongside instruction in the use of LexiURL searcher and link analysis methodology. In particular, an inductive method of categorisation will be used, and a random sample of websites identified will be analysed. This will attempt to illuminate the quantitative data gathered. However, use of content analysis involves other validity considerations.

Validity in relation to content analysis is taken to mean consistency between the categorisation of the data. Validity in webometric content analysis has been mixed. Some studies have shown high levels of validity between categories (Vaughan et. al. 2006), but the inconsistencies inherent in individuals' judgements in creating categories

and categorising results within them have been shown to create difficulties in other studies (Harries et. al. 2004). The validity of the content analysis in this investigation cannot be known without the use of additional researchers to categorise additional portions of the data gathered. Whilst this would have enabled a greater content analysis sample, and therefore increased the reliability of the investigations conclusions, such an approach with the resources available was not possible.

4. Findings and Discussion

4.1 Introduction

The aim of this chapter is to present the findings of the investigation through the analysis of the data gathered. This is in order to relate the findings to answering the research questions, in line with the research aims (Ryan 2006b). The aim of the research is to investigate the users of UK based institutional repositories through the use of webometric link analysis in identifying user groups and comparatively evaluating institutional repositories. The research questions that are derived from the research aims are:

- Can link analysis be used to identify repository user groups?
- Can link analysis be used as a comparative evaluation tool for institutional repositories?

The data gathered takes the form of quantitative findings, derived from the LexiURL searcher link analysis, and qualitative findings, where that data is subject to content analysis to give further insight. The chapter is introduced by setting out the sources of the data by giving descriptions of the institutions investigated, which have been presented anonymously to avoid ethical implications. The quantitative link analysis data is presented first, followed by the qualitative content analysis data. The results of the investigation are outlined in relation to the research questions and literature review, and the implications discussed.

The complete results of the pilot study, including the link analysis and content analysis data, is attached in the appendices as an example of the research instruments.

4.2 Institutions Investigated

In order to avoid any ethical implications in dealing with public institutions, it was decided to present the institutions investigated anonymously. This also avoided the need to seek permission from the institutions studied. In discussing the results, institutions will be referred to by letter, and descriptions of the institutions are set out below for comparison. Institutional information is taken in part from HERO (2009), HESA (2006) and the Guardian (2008), as well as specific university websites. Information regarding the institution's repository is taken from ROAR (Brody 2007), OpenDOAR (University of Nottingham 2008b), CSIC's (2009) web visibility ranking, and individual institutional repository websites. In describing the 'size' of institutions, total number of students is

used as a proxy measure of relative size.

Institution A is a medium sized, research focused university, ranked in the top 10 research institutions in the UK and within the top 100 universities internationally. The institutional repository was one of the earliest established in the UK, registered with ROAR in 2003. Consequently it is one of the largest institutional repositories in the UK, is one of the most highly ranked UK institutional repositories by web visibility, and has previously been subject to webometric investigation. It is based on the Eprints platform. The results of the investigation of this institutional repository were generated as part of the pilot study, which resulted in slightly fewer pages being analysed for content. The data was gathered by LexiURL on 7th January 2009.

Institution B is a smaller, medium sized university, with a strong international focus. The institution is ranked in the top 10 in the UK and top 10 internationally (THES). The institutional repository was also established relatively early, having been registered with ROAR in 2004. It is one of the largest repositories in the UK by number of items. It is also ranked within the top 100 institutional repositories worldwide by web visibility, placing it amongst the most visible UK institutional repositories. It is also based on the EPrints platform. The data was gathered by LexiURL on 14th February 2009.

Institution C is the largest institution in this study, and one of the oldest in the UK. The institutional repository is based on the DSpace software platform. The institutional repository was also registered with ROAR from 2004. In addition, the institutional repository contains a large number of records pertaining to widely known subject specific dataset, making it one of the largest repositories by number of items. The repository is ranked within the top 150 worldwide by web visibility. The data was gathered by LexiURL on 14th February 2009.

Institution D is the smallest institution in this study, and also one of the oldest in the UK. The institutional repository is based on the Fedora software platform, and was registered with ROAR from 2007. The repository was previously available on the Eprints platform from 2004, and is in the process of migrating all items. The repository was not ranked in the CSIC web visibility top 300 ranking, but is ranked within the top 20 UK institutional repositories by number of items. The data was gathered by LexiURL on 5th March 2009.

The institutions were selected according to the method set out in the research design chapter. This meant that the criteria for selection was a combination of the age of the repository, the size of the repository measured by the number of item records

maintained, and web visibility. These criteria were used to ensure that sufficiently well established repositories where investigated in order to produce good results. In practice the criteria produced a similar profile, i.e. the longer a repository has been established for, the more likely it is to have larger numbers of item records, and be well established on the web. In addition, however, the criteria also led to institutions with similar profiles being selected. The institutions selected for investigation were of similar sizes, with student numbers ranging from 20,000 to 25,000 (HESA 2006) and are all members of the Russell Group of Universities (Russell Group 2009), an association of research-focused UK universities. The fact that they are research focused may be reflected by having well established institutional repositories. The selection approach may have had an unintended bias for well established repositories affiliated to research intensive institutions with similar profiles. The institution's profile was not known or taken into consideration when they were selected for study, so the similarity between the institutions is coincidental.

The main factor limiting the extent of the investigation was time. This was the case for both the number of institutional repositories selected for investigation, and the number of web pages selected for content analysis. The number of links from web pages needed to perform a valid content analysis was in part determined with reference to Thelwall (2003), who suggests that around 40 links/pages are suitable for exploratory investigations.

4.3 Link Analysis Results

As described in the research methods chapter, the LexiURL searcher software queries the selected search engine and returns the results in the form of a link impact report. This contains an overview of search results, giving the total number of hits estimated by the search engine, and the number of actual hits returned by the search engine to LexiURL searcher. The results are analogous to the figures given when manually searching the web. The difference between the two figures has been discussed previously in the literature review chapter, and is mentioned again below.

The report also lists all pages containing links returned by the search engine. This list is then analysed in order to highlight trends in the data. The results of the analysis are presented here. The final part of the report is a random sample of the individual web pages returned by the search engine, for the purpose of content analysis or other further investigation. These are discussed further in the section dealing with the content analysis data.

4.3.1 Overview

As discussed in the research design chapter, the first part of the link impact report generated by LexiURL searcher is a summary of the results returned by the search engine.

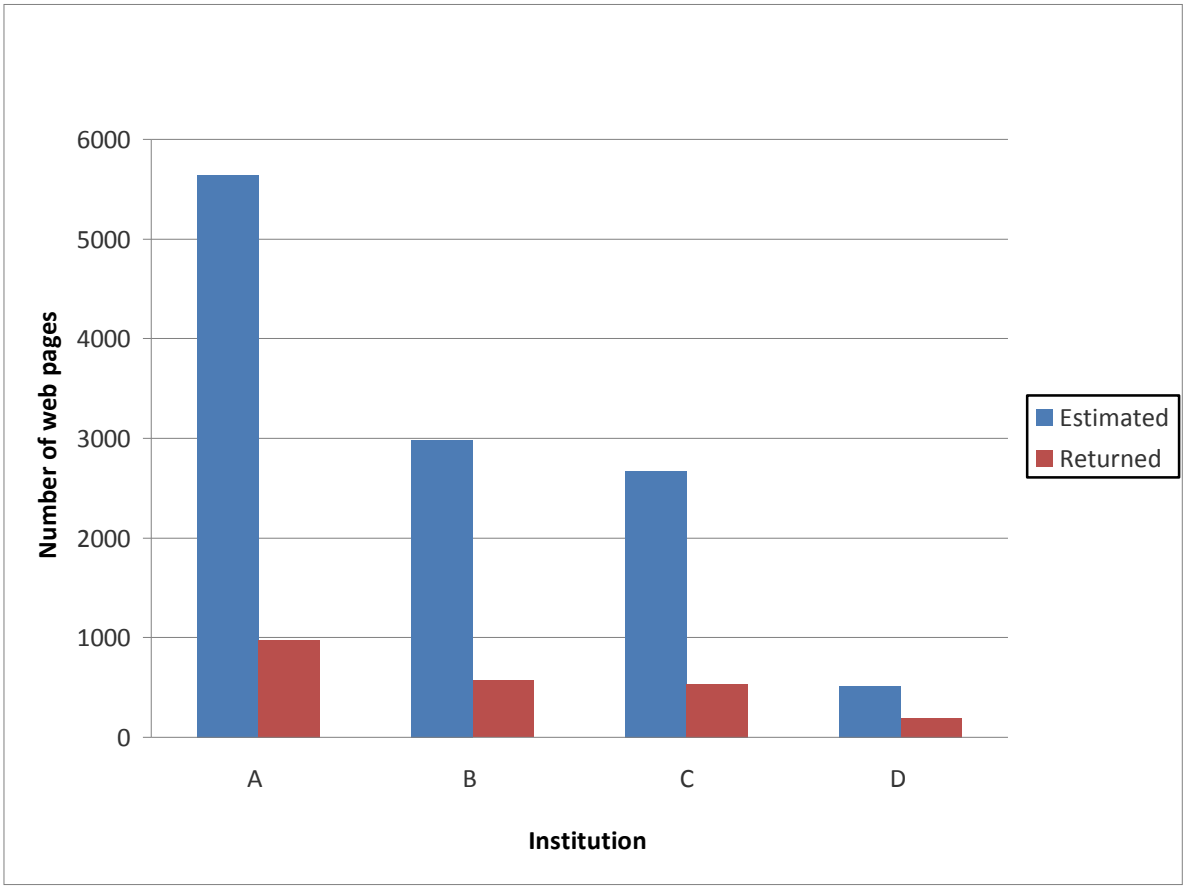

Figure 1. Histogram illustrating the number of estimated and returned web pages with links to each institutional repository.

As shown in the figure 1. LexiURL returned a wide range of values for the estimated total number of web pages containing links, and web pages containing links returned by the search engine, for each institutional repository. The institution with the highest number of results of both types was A, the longest established institutional repository, whereas the smallest number of results of both types was reported for institution D. As mentioned above, the institutional repository of D is currently in the process of being migrated from one software platform to another. This could explain why so few links to the website were retrieved in comparison with its peers. Fewer results are returned for institution C in comparison with A and B, despite C being the largest institution (by student population), and the largest repository by number of items. The number of results returned therefore seems to be more correlated with the age of the institutional

repository at its current web address, rather than characteristics of the home institution or the number of items held. This indicates that institutional repositories are indeed a web-embedded technology, as assumed in the research methods chapter. It also indicates that it would be difficult to give a comparative evaluation of the repositories based solely on the number of pages linking to the repository, as it only appears to be associated with the length of time the repository has been established.

It should be noted that the estimated number of hits returned by a search engine has been shown to be unreliable (Thelwall 2008a), and several reasons for the difference in the number of estimated hits and the number of hits returned have been put forward, mostly due to programmes and algorithms associated with the functioning of the search engine (Thelwall 2008b).

4.3.2 Top Level Domains

Figure 2 shows a graph of the pages linked to the four institutional repositories, summarised by TLD. The reason for summarising in this way is to show how links to the institutional repositories are distributed across the different domains of the web. The largest number of links is from the generic TLD .com, used by many different organisations. The second largest number of links is from the .uk domain. This is usually used by websites registered or affiliated with the UK. This will include the STLDs discussed below. The prevalence of the .uk TLD should reflect the fact the repositories investigated are UK based, and so have a mainly UK audience. In order to provide finer detail, figure three shows a breakdown of the .uk TLD into the STLDs reported by LexiURL searcher.

Other TLDs of note are .edu, assigned to American universities, .gov, assigned to American government departments. Country-specific TLDs reported for all repositories include Germany (.de), China (.cn) and Canada (.ca). In addition, a large number of country specific TLDs with small numbers of links are reported. This might indicate that repositories are finding a wider audience around the world, particularly in English speaking countries (USA and Canada). The distribution of TLDs reported by LexiURL searcher is similar between the repositories studied, and indicates that institutional repositories are attracting a similar TLD link profile. Because of the small numbers of links from country-specific TLDs outside the USA and UK it is difficult to see how this data could be used for comparative analysis of international impact. That is, not enough data is currently available to be able to adequately compare the repositories in this aspect of the investigation.

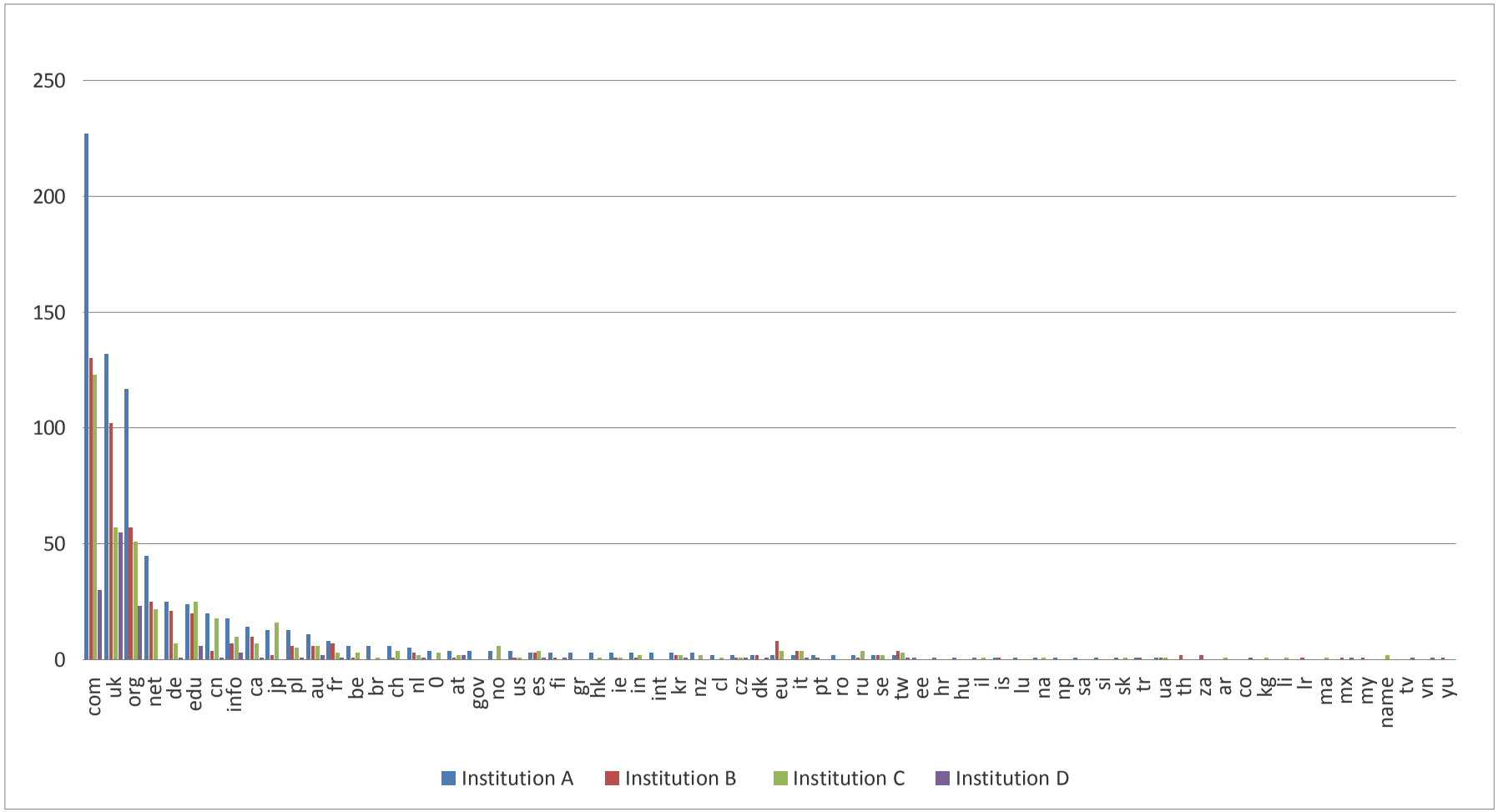

Figure 2. Histogram showing a summary of links to the institutional repositories, arranged by Top Level Domain.

4.3.3 Second Top Level Domains

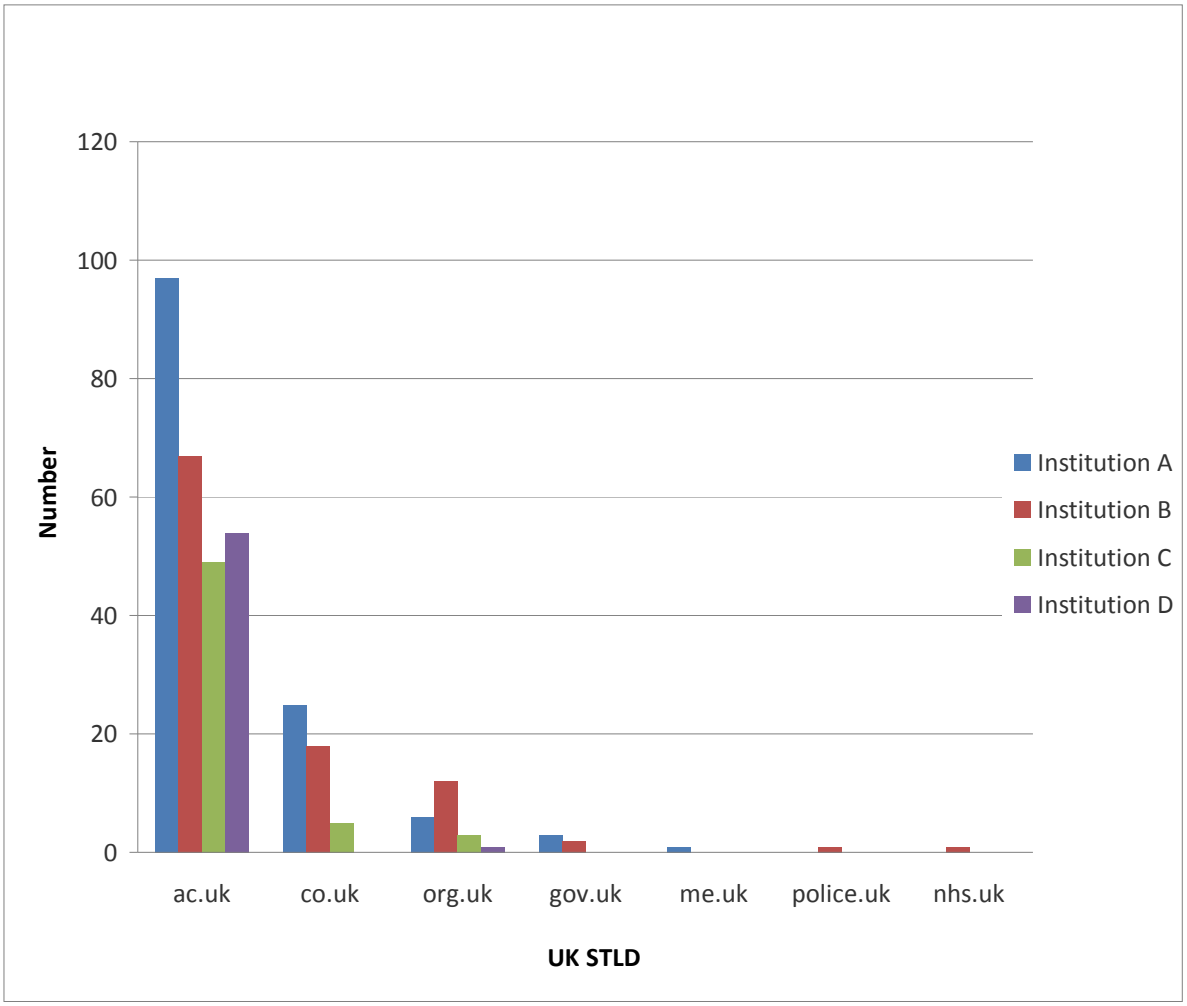

Figure 3. Histogram showing links to institutional repositories from UK STLDs

The graph shown in figure 3 shows the UK related STLDs reported by LexiURL searcher for the repositories studied. These results have been presented in isolation to show how UK repositories are embedded with other UK-based websites. It shows that the majority of links come from the academic-related STLD (.ac.uk), whilst most of the rest come from the generic STLD .co.uk. A small number of links come from STLDs related to government (.gov.uk) and other public services, police (.police.uk) and hospitals (.nhs.uk). This suggests that within the UK related web, institutional repositories are well embedded in the academic sector, with strong links from general websites, possibly including public, personal and commercial websites, and with some links to other public sector websites. The numbers of links to each repository reported follow the same patterns identified for the total number of links returned, discussed above. Even though describing the links to repositories by STLD gives a finer level of detail regarding types of website and user group, it is still not detailed enough to answer the research question regarding identifying user groups. A rough comparative analysis by STLD could be

undertaken if the information required was, for example, to check that similar numbers of links were originating from UK academic sites. Again, however, not enough information is contained in the STLD to make detailed comparative analysis possible.

4.4 Content Analysis

The second part of the investigation, as set out in the research design chapter, is that of the content analysis.

The random sample of in-links retrieved by LexiURL for each institutional repository investigated was divided into four categories for the purpose of the content analysis. These categories were based on the existing literature identified in the literature search, particularly Thelwall (2003) and Bar-Ilan (2005), as well attempting to answer the research questions. The four categories used are: the types of pages identified as containing links to the target website (in-links) or the Source Page; the pages identified as the target of in-links or Target Pages; the possible motivations for creating the identified links or in-link Motivation; what the source page, target page and link motivations reveal about the types of users likely to find or follow such a link, or Possible User Groups. The reasoning for organising the content analysis in this way, and discussion of the sub-categories used is laid out below.

4.4.1 General Observations

As discussed in the research design chapter in relation to the pilot study, there were a number of foreign language websites returned in the random sample by LexiURL searcher. In combination with the discussion above on TLDs, this indicates that institutional repositories are having a global impact through their free availability on the web. Websites in a non-English language were not noted for the pilot study results content analysis categorisation (institution A), but were noted for the subsequent repositories, as discussed in the research design chapter.

As the coverage of web search engines extends beyond links within web pages to include HTML links embedded in other types of documents available on the web, there were some instances of these documents appearing in the categorising exercise. These included Adobe Acrobat documents, and Microsoft Word, Excel and PowerPoint documents. These were treated as analogous to web pages for the purposes of the content analysis, and included in the categorisation as they constituted documents available on the web with HTML links to the target website. This meant they illustrated possible user groups and were indexed by web search engines, making them visible to

investigation by link analysis methods.

Some of the results recovered appeared to be randomly generated pages containing nonsense text and multiple random links, including those pointing to the institutional repositories under investigation. These types of sites are often referred to as link spam, and are usually produced automatically in an attempt to influence search engine results (Gyongyi and Garcia-Molina 2005). As search engine providers attempt to counter these sites, they must give the impression that they are legitimate sites, and so contain random or appropriated text, and links to a variety of legitimate web pages. They were treated as similar to foreign language websites for the purpose of the categorisation exercise, in that they contribute to the web visibility of the target institutional repository, but do not give any information regarding users or user groups. Their inclusion is in partial contrast with Thelwall's (2004) conditions for inclusion of links in a link analysis study, that links be created:

- Individually and independently
- By humans
- Through equivalent judgements about the quality of the information on the target page.

However, Thelwall is also realistic enough to note that these conditions are rarely met in full, partly because of the nature of the web, and partly because of human nature. These drawbacks are also used in part by Thelwall (2006) to explain why statistical analysis is not appropriate in small-scale link analysis investigations, and is not used here.

4.4.2 Types of pages containing in-links (Source pages)

The types of pages that contained links pointing to the institutional repository were recorded to see what types of source web pages contained links to institutional repository target pages. Similar categories to those used are found elsewhere in the literature (Vaughan et. al. 2007). They are defined here to make any assumptions explicit. The pages were initially sorted into broad categories, such as forum, blog etc. and then sub-divided into more specific categories where necessary, to distinguish, for example, between an academic homepage at the same institution as the repository, and one based at a different institution (Ryan 2006b). This process was in common with other studies identified (Vaughan et al. 2006, Bar-Ilan 2005), although this can cause ambiguity in categories. This was overcome in this investigation by categorising the different ‘types’ of web page in two stages. ‘Plain’ websites were classified according to their affiliation. ‘Other’ types of website were classified according to their type or

structure. Tables 1 and 2 set out the categories used with a description of each.

Table 1. Table showing 'plain' web page categories, descriptions and examples.

Category	Description
individual academic's publication page - internal	Any page containing a list of publications by an individual affiliated with the institution maintaining the repository
individual academic's publication page - external	Any page containing a list of publications by an individual not affiliated with the institution
departmental/research group page - internal	Any page produced by a department, research group or similar, affiliated with the institution
departmental/research group page - external	Any page produced by a department, research group or similar, not affiliated with the institution
government related	Any page produced by governmental body or similar, including research councils.
commercial/industrial related	Any page produced by a commercial entity
library or repository related	Any page produced by or related to a library service, repository service or similar

Table 2. Table showing 'other' web page categories, descriptions and examples.

Category	Description
non-html page	Any non-html document retrieved, for example PowerPoint slides
foreign language	Any page presented in a non-English language
blog	Any page presented in a weblog format
forum/discussion board	Any page presented in a forum or discussion board format
email list archive	Any page that presents an archived email list
wiki	Any page presented in a wiki format
social network page	Any page presented as part of a social network
other public information page	Any page presented by a public group, for example an NGO or private individual
automatically generated page	Any page generated without direct human input, for example, lists of search results, or link spam

Table 3. Table showing the categorisation of links to institutional repositories by type of source page.

Type of Source Page	Insititution A	Institution B	Institution C	Institution D
Staff Publications Page - Same Institution	2	3	1	1
Staff Publications Page - Different Institution	6	0	2	4
Departmental/ Research Group Page - Same Institution	1	5	3	6
Departmental/ research group page - Other/ cross Institution	1	4	4	5
Government-related webpage	2	1	0	1
Commercial/Industrial-related webpage	1	1	0	0
Foreign Language Page		5	20	6
Automatically generated page/link spam	9	4	2	3
Blog	3	2	5	3
Forum/Discussion Board	2	5	1	0
Email List Archive	3	0	2	1
Library/Repository related website - Same Institution	0	1	1	3
Library/Repository related website - Other institution	6	5	3	3
Other Public webpage	2	3	1	0
Wiki	1	2	9	0
Non-html page	0	2	2	4
Social Network or similar	0	1	1	2
Total number of source pages	39	44	57	42

Table 3 shows the random sample of source pages for the four institutional repositories, categorised as according to the above criteria. Presented in this way, it is possible to draw comparisons and discern patterns between the institutional repositories.

Academic-related pages were perhaps unsurprisingly a large proportion of the results

analysed. However, for some of the repositories there appeared to be more academic websites not directly affiliated with the home institution. This did include academics who had previously been members of the institution, and cross-institution research groups that included some members of the repository-related institution. The proportion of institution-affiliated academic publication lists reported were small, perhaps highlighting issues surrounding difficulties in engaging faculty members in depositing in repositories (Davis and Conolly 2007).

There were a number of more informal types of website contained in the samples, including newer types of web pages like social networks, blogs and wikis. These informal websites suggest that a wider range of web users are linking to, and hence using, institutional repositories. This could also indicate that academics are using services outside the institution as informal methods of scholarly communication, including blogs and social networks.

The inclusion of archived email lists in the results for most of the repositories investigated suggests that cybermetric techniques could also be usefully used in investigating repositories. This might identify user groups not found through webometric techniques.

Also included in the results sample were automatically generated pages, mentioned above. More of these pages were associated with institution A than the others investigated. This could again be a function of the age and length of time that institution A's repository has been established for, suggesting that although automatically generated pages are not useful for this study, and are generally considered to be a 'bad thing', they may form a proxy measure of how well a website is embedded within the web. However, automatically generated pages were found in association with all the repositories, and may just be an unfortunate side-effect of being a publicly accessible website.

The categories used in this part of the content analysis begin to give an idea of the types of user groups that might be associated with institutional repositories, including large numbers of academic-related users, and smaller but significant numbers of library-related, government and commercial users. However, contextual information regarding link motivation was needed to classify types of user groups in more detail.

4.4.3 Specificity of in-links (Target pages)

Previous studies (Bar-Ilan 2005, Vaughan et. al. 2006) have included the target page of

links (the page a link points to) as part of the investigation. This is in order to better understand the relationship between the 'source' page and the 'target' page when classifying link creation motivations. In this investigation, all the target pages were contained in the institutional repository, and so the institutional repository page 'type' would be constrained by the architecture of the institutional repository website. However, some interesting differences were noted between the institutional repositories as to the frequency of types of target pages. It is worth noting that more than one link of the same type was only recorded once, where as more than one link of a different type was recorded separately.

The categories used in the classification are laid out in table 4.

Table 4. Table showing categories for target page classification.

Category	Description
Homepage	a link to the root page of the institutional repository
Item page	a link to the page describing the item and its location.
Item	a link to the item itself (e.g. a pdf file)
Communities page	a link to a page describing a collection of items grouped as a community, for example a department or subject area
Collection page	a link to a page describing a collection of items grouped as a collection, for example a subject area or project
Technical page	a link to a page describing or containing technical information regarding the institutional repository, for example information relating to metadata harvesting.

Table 5. Random sample of links to institutional repositories categorised by target page type.

Target page type	Insititution A	Institution B	Institution C	Institution D
homepage	8	7	11	21
item	4	9	4	0
item page	24	20	12	14
collection page	0	0	6	0
community page	0	0	2	0
technical page	4	2	1	2
support page	0	2	0	0
broken link	0	1	2	0
other	0	0	0	1

Table 5 shows the results for the categorisation of the random sample of websites retrieved, arranged by the target of the link to the institutional repository. The table illustrates that for institution A and B, the highest number of links are directed at item pages, which usually contain a description, metadata and, where available, a link to the item itself. For institution C, a roughly equal number of links were found to item pages and the repository home page, whilst for institution D the majority of links pointed to the homepage. This again could be related to the relative lengths of time the repositories have been established for. As the repository becomes better populated, the proportion of navigational-related links to the homepage falls against the number of links to item pages that are of sufficient interest to be linked to directly.

All the repositories were found to have links to pages classified as technical. This is likely to be in part a reflection of a general trend identified of repository creators and maintainers (i.e. people and organisations involved in repositories, libraries or open access) creating links to repositories. However, this category also includes links to dynamic pages within repositories, created to showcase or publicise particular repository content, for example RSS feed pages or search results pages listing items by department. This was particularly found in relation to institution A, which may reflect its position as an early adopter of the technology.

It is not necessarily surprising that more links to item pages than items were found. Many repositories have more item pages listed than items that are available to download, as some authors are reluctant to make the full text of their articles available (Davis and Connolly 2007). In addition, it may be that item pages are more persistent in their web location than individual items, leading to more stable links over time, although there is no evidence to support this.

4.4.4 In-link motivations

The categories used to group link motivations are based in part on Thelwall (2003), who describes a number of categories of discerned link motivations, including ownership, social, navigational and gratuitous. Other classifications of link motivations have been undertaken that influence these categories, including Bar-Ilan (2005), Vaughan et. al. (2006) and Wilkinson et. al. (2003). In addition, the range of categories found in the literature indicates that it is acceptable to generate categories to suit the data. The understanding of what is meant by the category title is crucial, and the categories used here are defined below. As mentioned previously, links returned from automatically generated pages or foreign language pages were not included in this part of the analysis. In addition, websites in languages other than English were not included in the categorisation for link motivation as it would not be possible to infer the categories applicable without a working knowledge of the language. Table 6 defines the categories used in this analysis.

Table 6. Table showing categories for in-link motivation classification.

Category	Definition
Ownership	Used for both individual academic publication lists and departmental publication lists
Affiliation	Used to indicate a connection between the source and target pages that is collaborative or equal in nature, for example, linking between the repository and departments at the same institution
Recommendation	Although all links can be said to represent a recommendation to some degree (Zuccala et. al. 2006), this category was used when recommending sources for learning or research support
Reference	Not only in the strict academic sense, but including links for further information, citations from wikis and forums etc., and bookmark-type links
Responsibility	Used when claiming responsibility but not ownership, for example employees noting their relationship to the repository
Example	Used when a link is given as representative of a group or type

Table 7. Random sample of links to institutional repositories categorised by link motivations.

Link motivations	Insititution A	Institution B	Institution C	Institution D
ownership	10	11	3	12
affiliation	2	3	2	2
recommendation	4	10	5	5
reference	11	10	17	8
responsibility	0	0	6	5
example	3	1	5	4

The results of the categorisation of links returned by link motivations are shown in table 7. There are several interesting trends that can be identified in these results. Institutions A, B and D were found to have roughly equal numbers of links to the repository in the sample that were created to show ownership and reference. Institution C in contrast had many more links classified as reference than any other category. This may be in part due to the large subject-specific dataset held within the repository of institution C, garnering a large number of links through its usefulness in that subject, and hence influencing the data gathered. Overall, the links found in the sample of all four repositories suggests that the two most common reasons for linking to a repository are to show ownership of a resource, either as an individual or as a department or research group, and to provide a reference of some kind, including formal citations within electronically presented articles, but also informal references, including links to further information in resources like Wikipedia. All the repositories had in-links created to describe them as examples of repositories. This is possibly an extension of the trend identified in the literature of describing and commenting on repository development. In addition, it correlates with the results presented in table 3, where source pages associated with libraries and related organisations were found linking to all of the institutions repositories. All the repositories had similar numbers of in-links described as affiliation, where there is a connection between the creators of the source page and the repository. These links were from a range of origins, including from pages within the same institution, and from supporting organisations, such as JISC.

4.4.5 Possible user groups

Developing user groupings required the most intuitive usage of content analysis categorisation. Individual judgement as to what constitutes a user group and which group a website indicates was used frequently as no guiding principles were revealed in the literature review. Some groupings were more apparent than others, and guidance for creating user groupings was taken from the literature reviewed, including McKay (2007), Zuccala (2006). The classification of user groups was inductively created by looking at

the type of source website as a whole, the link motivation, and link context on the source page to identify what type of user might be represented by the link. There can be considerable overlap between the categories used in this classification. For example, a library web page recommending an institutional repository will be primarily for the benefit of teaching and research support, and so classified as academic-related, but may also represent use of the repository by library-related users in identifying suitable resources. In addition, it would be possible to argue that there is some overlap between the categories used in the aspects of the content analysis. For example, academic's homepages will be highly correlated with ownership link motivations, and be related to academic user groupings. The arrangement of the content analysis in this manner gives some structure to the categories that will ultimately be used to answer the research question regarding institutional repository users, and provides some justification for the categories used, and therefore the conclusions drawn. As mentioned above, websites in languages other than English were not included in the categorisation for possible user groups, as it would not be possible to infer the categories applicable without a working knowledge of the language.

The user groupings are defined in table 8 for clarity in discussing the results.

Table 8. Table showing categories for possible user group classification.

Category	Definition
Academic – Same Institution	Pages representing use by academics from the same institution as the repository
Academic – Cross Institution	Pages representing use by academics from other institutions
Academic support	Including the academic-related areas of teaching and research support
Open Access, Institutional Repository or Library and Information Science related	Pages intended for use by users related to repository maintenance or similar
Public	Pages created by and for use by public groups
Professional	Websites which are not individually focused, but not easily classified as academic, public or library related, including governmental, commercial and charity groups

Table 9. Random sample of links to institutional repositories categorised by possible user group types.

User group types	Insititution A	Institution B	Institution C	Institution D
Academic-same institution	3	8	5	4
Academic-different institution	12	3	7	9
Academic-support	2	6	7	7
Public	5	9	7	1
Repository related	7	9	11	15
Professional	3	8	0	0

The classification of the results sample by potential user group is shown in table 9. As before, the results are not surprising in showing the majority of users to be academic related. The results are able to show that currently similar numbers of links to repositories are related to the administration of repositories as are related to academic use of repositories. This is perhaps a reflection of the early stage of institutional repository deployment in the UK, and will change over time as repositories become more embedded in scholarly communication workflows.

The pattern seen in table 3 regarding academic use of the repository within and outside of the institution is repeated in table 9, with institutions A, C and D having more academic users from outside of the institution, and B having more users within. Institution B has the most public- and professional-related links, suggesting that it has the most widely used repository outside of academia, but institutions C and D have the most academic support-related links, suggesting that these repositories are used more extensively in teaching and research support.

These results could be used for a qualitative comparative evaluation between the repositories in this investigation. The results suggest that each repository investigated has user groups that are more likely to indicate usage of the repository by creating in-links. By examining the types of links to peer repositories, administrators can attempt improving usage from similar user groups.

Although the categories and results are similar between table 3 and table 9 there is a difference in how the pages were classified, in particular, general pages, such as forums, blogs and email lists were assigned to user groups. Also, sample pages could be assigned to more than one user group where this was felt to be appropriate. However, the level of detail in distinguishing between user groups that was hoped for was not

achieved. It was found that fine levels of distinction when classifying user groups based on links from individual web pages was difficult, due to the individual nature of motivations for link creation. Because of this difficulty, the results are only able to indicate broad types of user group. If a more detailed level of information regarding users was needed, it would be useful to have a larger sample size.

5. Conclusions

5.1 Introduction

This chapter concludes the research into institutional repository user groups through link analysis. It discusses the results in relation to the research questions, summarises the drawbacks identified in the methodology and its application, discusses the implications of the results for professional practice and provides suggestions for further study based on this research.

5.2 Answering the Research Questions

The aim of the research project was to investigate the types of users of institutional repositories and the use of link analysis as a tool to reveal such groups. In addition, the investigation aimed to look at the possible use of link analysis as a comparative evaluative tool for institutional repositories. The results illustrated that link analysis in combination with content analysis was an investigative tool that could be used to give an indication of types of user groups as a proxy measure of individual users. The results also indicated that for the institutions in this investigation, link analysis alone did not give a good comparative evaluation of institutional repositories. Content analysis in combination with link analysis can give a range of comparative evaluation metrics, but the issue of validity between categories is not addressed in this investigation, and the cost resources for a suitably comprehensive evaluation would likely be prohibitive.

5.2.1 Can link analysis be used to identify institutional repository user groups?

Link analysis data was gathered in the form of a link impact report, using the software LexiURL searcher. This data comprised lists of web pages identified as having links to the institutional repositories included in the investigation. On its own, the data is only able to give very broad information regarding user groups. Results collated by TLD allow some indication of international use of repositories. Sufficient results for collation by STLD were only available for UK based domains, and indicated that most links from UK web pages came from academic websites. Overall, it was found that there was insufficient detail in the link analysis data gathered to identify institutional repository user groups at a finer level of detail.

Link analysis has been found to be most effective when coupled with other investigative methods. In this investigation, content analysis was used to investigate a sample of web

pages containing links to repositories in greater detail. The content analysis was able to give very broad descriptions of institutional repository user groups, but for more detailed results a much larger sample would need to be classified.

5.2.2 Can link analysis be used as a comparative evaluation tool for institutional repositories?

Link analysis has been proposed as a comparative evaluative tool for website managers (Thelwall 2009a), and involved in instances of evaluation of digital repositories (Zuccala et. al. 2006, 2007, 2008). In this investigation, data was gathered through link analysis methods with the intention of comparing aspects of different repositories. It was found that the total number of links to a repository was more closely associated with the amount of time a repository had been established in the current web location, rather than other factors commonly used for comparative evaluations, such as number of items held, or profile of the associated institution. In looking at in-linking web pages arranged by TLD and STLD similar results were found. It would appear from these results that link analysis alone does not provide suitable data for comparative analysis.

5.3 Evaluation

The results as they are presented allow several useful conclusions to be drawn. However, it is important to remember that there are drawbacks to the methodology used in several areas, which have implications for the validity of the conclusions drawn.

Link analysis is a quantitative methodology used to investigate hyperlinks between web pages. Its application in this investigation has been reliant on the software LexiURL searcher. Whilst this is produced for investigative purposes by academics, there is very little evaluative literature regarding its outcomes, and no information available regarding its workings. In addition, the data was gathered via a web search engine. There are several problems with data gathered by search engines discussed in the research design chapter.

Content analysis is a qualitative methodology, and can be applied in many situations to try and determine meaning from text. In order to draw valid conclusions from content analysis it is recommended that multiple researchers classify the content investigated (Weber 1990). In addition, Thelwall (2009) notes that for an adequate sample of links between web pages the number of pages needed to be classified can be very large. Neither of these conditions was met in this investigation.

The results of the investigation should therefore not be taken as conclusive evidence of the trends identified in the findings and conclusions. Rather, as there is a noted lack of investigation into users of institutional repositories in the available literature (McKay 2007), this is an exploratory study which suggests and implements some possible methods for investigating the research aims and questions.

5.4 Implications

Institutional repositories are of continuing significance to UK higher education and associated library services. The literature review has illustrated that there is a lack of knowledge as to what types of user groups are associated with repositories. The outcome of this investigation, that illustrates a number of user groups associated with repositories, both confirms a number of assumptions regarding repository users, and also suggests possible target groups that could be included in strategic planning regarding future repository development. Repositories are mainly used by academic related user groups, although not necessarily formally and not usually associated with the institution hosting the repository. Linking for support of other academic processes, such as teaching, was found to be a relatively small part of overall academic user groups. Other user groups uncovered that may not be instinctively associated with institutional repositories include public users accessing repositories for information for non-academic reasons. There are also a number of professional user groups associated with repositories, including those using government and commercial websites. These user groups are important in widening access to academic research output, both formally and informally, which is important for institutional repositories and HE in general.

The use of link analysis in evaluating repositories is first made in Zuccala (2006). It is suggested that repository managers can undertake link analysis studies at six month intervals to check how and which users are linking to the repository. This author's experience of a link analysis study would suggest that a full link analysis study would be too time consuming for managers to repeat at such regular intervals. The CSIC ranking of repositories has shown that it is possible to use webometric methods to evaluate repositories. However, the results of this investigation have shown that the link analysis methods employed here are not suitable for fully evaluating institutional repositories.

5.5 Future Research

The research methods and conclusions suggest several ways that this investigation could be built upon or improved. The usual suggestion is to increase the amount of data

gathered. This could be achieved by looking at more repositories in the link analysis, or increasing the sample of websites in the content analysis. Variations of this would include comparing an international sample of repositories, to look for differences in user groups in different countries. Alternatively, a more exclusively ethnographic approach could be taken, and an exhaustive in-depth investigation of one repository could be undertaken, looking at all links to a single repository and categorising for user groups.

Webometric techniques, which include link analysis, have traditionally been benchmarked against other techniques in order to determine their validity. As there is very little parallel research currently into repository users, this is a technique that could not be directly applied at the present time. However, in order to give a better idea of validity, future research could compare the types of user profile associated with institutional repositories with other types of scholarly communication available freely on the web. This could include formal types, such as OA journals, and informal types such as scholarly blogs and homepages. In addition, to confirm the validity of the conclusions in this investigation, future research should focus on alternative techniques that could be used to validate this investigation's results. In particular, there are several alternative methods mentioned in the research design chapter that can be used in combination with link analysis if the resources were available. These comprise webometric techniques, such as web log analysis, and more traditional techniques such as surveys and focus groups. These techniques can be used in combination with link analysis to provide complimentary data to support the conclusions drawn through link analysis.

6. References and Bibliography

Adams, A. (2007) Copyright and research: an archivangelist's perspective. *SCRIPTed* [online] 4(3). Available via: http://www.law.ed.ac.uk/ahrc/script-ed/vol4-3/adams.asp [accessed 26 March 2009].

Allard, S., Mack, T. and Feltner-Reichert, M. (2005) The librarian's role in institutional repositories: A content analysis of the literature. *Reference Services Review* [online] 33 (3). Available via: http://dx.doi.org/10.1108/00907320510611357 [accessed 19 April 2009].

Almind, T. and Ingwersen, P. (1997) Informetric analyses on the world wide web: methodological approaches to 'webometrics'. *Journal of Documentation* [online] 53 (4). Available via: http://dx.doi.org/10.1108/eum0000000007205 [accessed 19 April 2009].

Andrews, R. (2003) *Research Questions*. Continuum Research Methods. London: Continuum.

Armbruster, C. (2008) *Usage and Citation Metrics: What Function for Digital Libraries and Repositories in Research Evaluation?* [online]. Available via: http://ssrn.com/abstract=1088453 [accessed 19 April 2009].

arXiv.org (2009) *General Information About arXiv* [online]. Available via: http://arxiv.org/help/general [accessed 25 January 2009].

Aschenbrenner, A. et al. (2008) The Future of Repositories? Patterns for (Cross-) Repository Architectures. *D-lib Magazine* [online] 14 (11/12). Available via: http://dx.doi.org/10.1045/november2008-aschenbrenner [accessed 19 April 2009].

Bar-Ilan, J. (2005) What do we know about links and linking? A framework for studying links in academic environments. *Information Processing and Management* [online] 41(4). Available via: http://dx.doi.org/10.1016/j.ipm.2004.02.005 [accessed 26 January 2009].

Bar-Ilan, J. (2008) Informetrics at the beginning of the 21st century—A review. *Journal of Informetrics* [online] 2(1). Available via: http://dx.doi.org/10.1016/j.joi.2007.11.001 [accessed 3 April 2009].

Barton, M. and Waters, M. (2004) *Creating an Institutional Repository: LEADIRS Workbook* [online] Available via: http://dspace.mit.edu/handle/1721.1/26698?show=full

[accessed 19 April 2009].

Barwick, J. (2007) Building an institutional repository at Loughborough University: some experiences. *Program:electronic library and information systems*. [online] Available via: http://dx.doi.org/10.1108/00330330710742890 [accessed 15 October 2008].

Bates, L. E. (2008) *"It's up by the six-fifties": investigating customer wayfinding in an academic library* [online]. MSc dissertation. University of the West of England. Available via: http://library.uwe.ac.uk/Archimages/31234.PDF [accessed 23 April 2009].

Beaulieu, A. (2005) Sociable Hyperlinks: an Ethnographic Approach to Connectivity. *In:* Hine, C. ed. *Virtual Methods: Issues in Social Research on the Internet.* Oxford: Berg.

Belden, D. M. (2006) *Weaving a Web of Precious Materials: Hyperlinks to, from and between Some Special Collections Libraries* [online]. MSLS dissertation. University of North Carolina. Available via: http://hdl.handle.net/1901/335 [accessed 15 February 2009].

Bevan, S. (2007) Developing an institutional repository: Cranfield QUEprints - a case study. *OCLC Systems and Services*. [online]. Available via: http://dx.doi.org/10.1108/10650750710748478 [accessed 14 October 2008].

Bergman, M. K. (2001) The deep Web: Surfacing hidden value. *Journal of Electronic Publishing* [online] 7(1). Available via: http://dx.doi.org/10.3998/3336451.0007.104 [accessed 20 November 2008].

Berkley Electronic Press (2008) *Digital Commons* [online]. Available via: http://www.bepress.com/ir/ [accessed 3 December 2008].

Björneborn, L. (2004) *Small-world link structures across an academic Web space: A library and information science approach* [online]. PhD Thesis. Royal School of Library and Information Science, Copenhagen, Denmark. Available via: http://vip.db.dk/lb/phd/ [accessed 20 November 2008].

Björneborn, L. & Ingwersen, P. (2004) Toward a basic framework for webometrics. *Journal of the American Society for Information Science & Technology* [online] 55(14). Available via: http://www.db.dk/binaries/PerspectivesWebometrics-Jasist.pdf [accessed 25 November 2008].

Blaxter, L. (2006) *How to Research*. 3rd ed. Maidenhead: Open University Press.

Branin, J. (2005) Institutional Repositories. *In:* Drake, M. A. ed. *Encyclopedia of Library and Information Science*. 2nd ed. New York: Marcel Dekker Inc.

Brin, S., & Page, L. (1998) The anatomy of a large scale hypertextual Web search engine. *Computer Networks and ISDN Systems* [online] 30(1-7). Available via: http://dx.doi.org/10.1016/S0169-7552(98)00110-X [accessed 17 January 2009].

Brody, T. (2007) *Registry of Open Access Repositories (ROAR)* [online]. Available via: http://roar.eprints.org/ [accessed 22 April 2009].

Brophy, P. (2005) *The academic library.* 2nd ed. London: Facet.

Budapest Open Access Initiative (2004) *Open Access Journal Business Guides* [online]. Available via: http://www.soros.org/openaccess/oajguides/ [accessed 20 December 2008].

Buehler, M. and Trauernicht, M (2007) From digital library to institutional repository: a brief look at one library's path. *OCLC Systems & Services* [online] 23(4). Available via: http://dx.doi.org/10.1108/10650750710831529 [accessed 19 April 2009].

Campbell, J., Daft, R. and Hulin, C. (1987) *What to study : generating and developing research questions.* London: Sage.

Carr, L., Brody, T. and Swan, A. (2008) Repository Statistics: What Do We Want to Know? *In: Third International Conference on Open Repositories 2008 Southampton, United Kingdom, 1-4 April 2008* [online]. Available via: http://pubs.or08.ecs.soton.ac.uk/30/ [accessed 10 February 2009].

Chan, L. (2004) Supporting and Enhancing Scholarship in the Digital Age: The Role of Open Access Institutional Repository. *Canadian Journal of Communication* [online] 29(3). Available via: http://www.cjc-online.ca/index.php/journal/article/viewArticle/1455/1579 [accessed 19 April 2009].

Cherry, J. and Duff, W. (2002) Studying digital library users over time: a follow-up survey of Early Canadiana Online. *Information Research* [online] 7(2). Available via: http://InformationR.net/ir/paper123.html [accessed 6 April 2009].

Creswell, J. (2003) *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*. 2nd ed. London: Sage.

Crow, R. (2002) The Case for Institutional Repositories: A SPARC Position Paper. *ARL*

Bimonthly Report [online] 223. Available via: http://www.arl.org/sparc/bm~doc/ir_final_release_102-2.pdf [accessed 25 September 2008].

CSIC (2009) *Ranking Web of World Repositories* [online]. Available via: http://repositories.webometrics.info/ [accessed 1 Feb 2009].

Davis, P. (2008) Author-choice open-access publishing in the biological and medical literature: A citation analysis. *Journal of the American Society for Information Science and Technology* [online] 60(1). Available via: http://dx.doi.org/10.1002/asi.20965 [accessed 10 February 2009].

Davis, P. and Connolly, M. (2007) Institutional Repositories: Evaluating the reasons for non-use of Cornell University's instillation of DSpace. *D-Lib Magazine* [online] 13(3/4). Available via: http://dx.doi.org/10.1045/march2007-davis [accessed 21 March 2009].

DSpace (2008) *DSpace Mailing Lists* [online]. Available via: http://www.dspace.org/feedback/mailing.html [accessed 3 January 2009].

EPrints (2008) *Welcome to the EPrints Wiki* [online]. Available via: http://wiki.eprints.org/w/Main_Page [accessed 3 January 2009].

Fidel, R. (2008) Are we there yet?: Mixed methods research in library and information science. *Library & Information Science Research* [online] 30(4). Available via: http://dx.doi.org/10.1016/j.lisr.2008.04.001 [accessed 20 December 2008].

Fuhr, N., Hansen, P., Mabe, M., Micsik, A. and Sølvberg, I. (2001) Digital Libraries: A Generic Classification and Evaluation Scheme. *In: Proceedings of the 5th European Conference on Research and Advanced Technology for Digital Libraries, Darmstadt, Germany September 4-9* [online]. Available via: http://dx.doi.org/10.1007/3-540-44796-2 [accessed 6 April 2009].

Fuhr, N. et. al. (2007) Evaluation of Digital Libraries. *International Journal on Digital Libraries* [online] 8(1). Available via: http://dx.doi.org/10.1007/s00799-007-0011-z [accessed 6 April 2009].

Govcom.org (200?) *Issuecrawler.net Scenarios of use for NGOs and other researchers* [online] Available via: http://www.govcom.org/scenarios_use.html [accessed 28 January 2009].

Guardian (2008) *Universities* [online] Available via: http://www.guardian.co.uk/education/list/educationinstitution [accessed 21 March 2009].

Gyongyi, Zoltan and Garcia-Molina, Hector (2005) Web Spam Taxonomy. *In: First International Workshop on Adversarial Information Retrieval on the Web (AIRWeb 2005), May 10-14, 2005, Chiba, Japan* [online]. Available via: http://ilpubs.stanford.edu:8090/771/ [accessed 7 March 2009].

Harnad, S. (1999) Free at Last: The Future of Peer-Reviewed Journals. *D-Lib Magazine* [online] 5(12). Available via: http://www.dlib.org/dlib/december99/12harnad.html [accessed 15 August 2008].

Harnad, S. (2001) The self-archiving initiative: Freeing the refereed research literature online. *Nature* [online] 410. Available via: http://www.nature.com/nature/debates/e-access/Articles/harnad.html [accessed 11 February 2009].

Harnad, S. and Brody, T. (2004) Comparing the Impact of Open Access (OA) vs. Non-OA Articles in the Same Journals. *D-Lib Magazine* [online] 10(6). Available via: http://www.dlib.org/dlib/june04/harnad/06harnad.html [accessed 19 April 2009].

Harries, G., Wilkinson, D., Price, E., Fairclough, R. and Thelwall, M. (2004). Hyperlinks as a data source for science mapping. *Journal of Information Science* [online] 30(5). Available via: http://dx.doi.org/10.1177/0165551504046736 [accessed 17 February 2009].

Harter, S. and Ford, C. (2000) Web-based analyses of E-journal impact: Approaches, problems, and issues. *Journal of the American Society for Information Science* [online] 51(13). Available via: http://www3.interscience.wiley.com/journal/73001123/abstract [accessed 7 March 2009].

HERO (2009) *Universities and College finder* [online]. Available via: http://www.hero.ac.uk/uk/universities___colleges/hei_listing.cfm [accessed 21 March 2009].

HESA (2006) *HESA Statistics – Higher Education numbers 2005/2006* [online]. Available via: http://www.hesa.ac.uk/dox/dataTables/studentsAndQualifiers/download/institution0506.xls [accessed 21 March 2009].

Hockx-Yu, H. (2006). Digital preservation in the context of institutional repositories. *Program: electronic library and information systems* [online] 40(3). Available via: http://dx.doi.org/10.1108/00330330610681312 [accessed 15 December 2008].

Houghton, J., Rasmussen, B., Sheehan, P., Oppenheim, C., Morris, A., Creaser, C., Greenwood, H., Summers, M. and Gourlay, Adrian R. (2009). *Economic implications of alternative scholarly publishing models : exploring the costs and benefits* [online]. Available via: http://www.jisc.ac.uk/publications/publications/economicpublishingmodelsfinalreport.aspx [accessed 2 February 2009].

Jankowski, N. and van Selm, M. (2005) Methodological Concerns and Innovations in Internet Research, *In:* Hine, C. ed. *Virtual Methods: Issues in Social Research on the Internet*. Oxford: Berg.

JISC (2005) *Opening up Access to Research Results Questions and Answers* [online]. Available via: http://www.jisc.ac.uk/uploaded_documents/QandA-Doc-final.pdf [accessed 19 April 2009].

Jones, R. (2006) Institutional Repositories *In:* K. Garnes, A. Landøy and A. Repanovici eds. *Aspects of Digital Libraries* [online]. Norway: Alvheim & Eide. Available via: https://bora.uib.no/handle/1956/1829 [accessed 19 April 2009].

Jones, R., Andrew, T. and MacColl, J. (2006) *The Institutional Repository.* Oxford: Chandos.

Kennan, M. and Wilson, C. (2006) Institutional repositories: review and an information systems perspective. *Library Management* [online] 27(4/5). Available via: http://dx.doi.org/10.1108/01435120610668179 [accessed 19 April 2009].

Kim, J. (2006) Finding Documents in a Digital Institutional Repository: DSpace and ePrints. *In: Proc. 68th Annual Meeting of the American Society for Information Science and Technology. Charlotte, North Carolina, 28 October - 2 November 2005* [online]. Available via: http://eprints.rclis.org/5189/ [accessed 27 May 2008].

Kim, H. H. and Kim Y. H. (2007) An Evaluation Model for the National Consortium of Institutional Repositories of Korean Universities. *Proceedings of the American Society for Information Science and Technology* [online] 43 (1). Available via: http://dx.doi.org/10.1002/meet.1450430176 [accessed 19 April 2009].

Kousha, K. (2005). Webometrics and Scholarly Communication: An Overview. *Quarterly Journal of the National Library of Iran* [online] 14(4). Available via: http://www.nlai.ir/Portals/2/files/faslname/60/en_article.pdf [accessed 5 January 2009].

LexiURL Searcher (2008) *LexiURL Searcher Web Analysis Software* [online]. Available via: http://lexiurl.wlv.ac.uk/ [accessed 9 December 2008].

Lynch, C. (2003) Institutional Repositories: Essential Infrastructure for Scholarship in the Digital Age. *ARL Bimonthly Report* [online] 226. Available via: http://www.arl.org/resources/pubs/br/br226/br226ir.shtml [accessed 19 April 2009].

Mark, T. and Shearer, M.K. (2006) Institutional Repositories: A Review of Content Management Strategies*, In: World Library and Information Congress: 72nd IFLA General Conference and Council. 2006*: Seoul, Korea [online]. Available via: http://www.ifla.org/IV/ifla72/papers/155-Mark_Shearer-en.pdf [accessed 26 July 2008].

Markland, M (2006) Institutional repositories in the UK: What can the Google user find there? *Journal of Librarianship and Information Science* [online] 38(4). Available via: http://dx.doi.org/10.1177/0961000606070587 [accessed 19 April 2009].

McCown, F., Liu, X., Nelson, M., and Zubair, M. (2006) Search Engine Coverage of the OAI-PMH Corpus. *IEEE Internet Computing* [online] 10(2). Available via: http://doi.ieeecomputersociety.org/10.1109/MIC.2006.41 [accessed 1 January 2009].

McGuigan, G. and Russell, R. (2008) The Business of Academic Publishing: A Strategic Analysis of the Academic Journal Publishing Industry and its Impact on the Future of Scholarly Publishing. *Electronic Journal of Academic and Special Librarianship* [online] 9(3). Available via: http://southernlibrarianship.icaap.org/content/v09n03/mcguigan_g01.html [accessed 3 January 2009].

McKay, D. (2007) Institutional Repositories and Their 'Other' Users: Usability Beyond Authors. *Ariadne* [online] 52. Available via: http://www.ariadne.ac.uk/issue52/mckay [accessed 26 May 2008].

Nicholas, D., Huntington, P., Jamali, H. R. and Tenopir C. (2006) Finding Information in (Very Large) Digital Libraries: A Deep Log Approach to Determining Differences in Use According to Method of Access. *The Journal of Academic Librarianship* [online] 32(2). Available via: http://dx.doi.org/10.1016/j.acalib.2005.12.005 [accessed 8 December 2008].

Nicholas, D., Huntington, P., Jamali, H. R. and Tenopir C. (2007) The impact of open access publishing (and other access initiatives) on use and users of digital scholarly journals. *Learned Publishing* 20(1).

Okerson, A., O'Donnell, J. (1995). *Scholarly Journals at the Crossroads: A subversive Proposal for Electronic Publishing* [online]. Available from: http://www.arl.org/scomm/subversive/ [accessed 20 December 2008].

Open Repository (2008) *Welcome to Open Repository* [online]. Available at: http://www.openrepository.com/ [Accessed 20 November 2008].

Organ, M. (2006) Download Statistics - What Do They Tell Us? *D-Lib Magazine* [online] 12(11). Available via: http://dx.doi.org/10.1045/november2006-organ [accessed 19 April 2009].

Orme, V. (2007) *"You will be..." A study of job advertisements to determine employers' requirements of library and information professionals in the UK* [online]. MSc dissertation. University of the West of England. Available via: http://library.uwe.ac.uk/Archimages/29477.PDF [accessed 19 April 2009].

Patton, M. Q. (2002). *Qualitative research and evaluation methods.* 3rd ed. Thousand Oaks, CA: Sage.

Payne, N. and Thelwall, M. (2007) A longitudinal study of academic webs: Growth and stabilisation. *Scientometrics* [online] 71(3). Available via: http://dx.doi.org/10.1007/s11192-007-1695-y [accessed 20 December 2008].

Pickton, M. J. (2006) *Research students and the Loughborough institutional repository* [online]. MSc dissertation. Loughborough University Available via: http://hdl.handle.net/2134/571 [accessed 23 August 2008].

Punch, K. (2006) Developing Effective Research Proposals. 2nd ed. London: Sage.

Ranganathan, S.R. (1931). *The Five Laws of Library Science* [online]. Available via: http://dlist.sir.arizona.edu/1220/ [Accessed 11 December 2008].

Robertson, R., Mahey, M. and Allinson, J. (2008) *An Ecological Approach To Repository And Service Interactions* [online]. Available via: http://ie-repository.jisc.ac.uk/272/ [accessed 21 January 2009].

Rowlands, I., Nicholas, D. and Huntington, P. (2004) Scholarly communication in the

digital environment: what do authors want? *Learned Publishing* [online] 17(4). Available via: http://dx.doi.org/10.1087/0953151042321680 [accessed 20 November 2008].

Rowlands, I. & Nicholas, D. (2005) Scholarly communication in the digital environment: The 2005 survey of journal author behaviour and attitudes. *Aslib Proceedings* [online] 57(6). Available via: http://www.emeraldinsight.com/10.1108/00012530510634226 [accessed 20 November 2008].

Rugg, G. (2007) *A Gentle Guide to Research Methods*. Maidenhead: Open University Press.

Russell Group (2009) *About the Russell Group* [online] Available via: http://www.russellgroup.ac.uk/ [accessed 11 March 2009].

Ryan, A. (2006a) Doing a Review of Literature. *In:* Ryan, A. ed. *Researching and Writing your Thesis: a guide for postgraduate students.* Republic of Ireland: Maynooth Adult and Community Education [online]. Available via: http://eprints.nuim.ie/873/ [accessed 9 February 2009].

Ryan, Anne B. (2006b) Methodology: Analysing Qualitative Data and Writing up your Findings. *In:* Ryan, A. ed. *Researching and Writing your Thesis: a guide for postgraduate students.* Republic of Ireland: Maynooth Adult and Community Education [online]. Available via: http://eprints.nuim.ie/871/ [accessed 10 February 2009].

Ryan, Anne and Walsh, Tony (2006) Constructing Your Thesis. *In:* Ryan, A. ed. *Researching and Writing your Thesis: a guide for postgraduate students.* Republic of Ireland: Maynooth Adult and Community Education [online]. Available via: http://eprints.nuim.ie/976/ [accessed 4 April 2009].

Salo, D. (2008a) *What do we want from IRs, and what are we doing to repository rats?* [online]. Available via: http://cavlec.yarinareth.net/2008/09/10/what-do-we-want-from-irs-and-what-are-we-doing-to-repository-rats/ [Accessed 19 November 2008].

Salo, D. (2008). Innkeeper at the Roach Motel. *Library Trends* [online] 57(2). Available via: http://digital.library.wisc.edu/1793/22088 [accessed 19 December 2008].

Schmitz, D. (2008) *The Seamless Cyberinfrastructure: The Challenges of Studying Users of Mass Digitization and Institutional Repositories* [online] Council on Library and Information Resources. Available via: http://www.clir.org/pubs/archives/schmitz.pdf [accessed 19 April 2009].

Seadle, M. (2008) The digital library in 100 years: damage control. *Library Hi Tech* [online] 26(1). Available via: http://dx.doi.org/10.1108/07378830810857744 [accessed 30 September 2008].

Statistical Cybermetrics Research Group (2008) *About the Statistical Cybermetrics Research Group* [online]. Available via: http://cybermetrics.wlv.ac.uk/about.html [accessed 30 December 2008].

Thelwall, M. (2001) Results from a web impact factor crawler. *Journal of Documentation* [online] 57(2). Available via: http://dx.doi.org/10.1108/EUM0000000007081 [accessed 7 March 2009].

Thelwall, M. (2002a) Methodologies for crawler based Web surveys *Internet Research* [online] 12(2). Available via: http://dx.doi.org/10.1108/10662240210422503 [accessed 19 April 2009].

Thelwall, M. (2002b) A comparison of sources of links for academic Web impact factor calculations. *Journal of Documentation* [online] 58(1). Available via: http://dx.doi.org/10.1108/00220410210425412 [accessed 19 April 2009].

Thelwall, M. (2003) What is this link doing here? Beginning a fine-grained process of identifying reasons for academic hyperlink creation. *Information Research* [online] 8(3). Available via: http://informationr.net/ir/8-3/paper151.html [accessed Feb 20 2009].

Thelwall, M. (2004) *Link analysis: An information science approach.* San Diego: Academic Press.

Thelwall, M. (2005a) Scientific web intelligence: finding relationships in university webs. *Communications of the ACM* [online] 48(7). Available via: http://dx.doi.org/10.1145/1070838.1070843 [accessed 19 April 2009].

Thelwall, M. (2005b). Webometrics. *In:* Drake, M. A. ed. *Encyclopedia of Library and Information Science* 2nd ed. New York: Marcel Dekker, Inc.

Thelwall, M. (2006) Interpreting social science link analysis research: A theoretical framework. *Journal of the American Society for Information Science and Technology* [online] 57(1). Available via: http://dx.doi.org/10.1002/asi.20253 [accessed 19 April 2009].

Thelwall, M. (2007a) Bibliometrics to Webometrics. *Journal of Information Science*

[online] 34(4). Available via: http://dx.doi.org/10.1177/0165551507087238 [accessed 19 April 2009].

Thelwall, M. (2007b) *Webometrics* [online]. Available via: http://webometrics.blogspot.com/ [accessed 17 February 2009].

Thelwall, M. (2008a). Extracting accurate and complete results from search engines: Case study Windows Live. *Journal of the American Society for Information Science and Technology* [online] 59(1). Available via: http://dx.doi.org/10.1002/asi.20704 [accessed 5 April 2009].

Thelwall, M. (2008b). Quantitative comparisons of search engine results. *Journal of the American Society for Information Science and Technology* [online] 59(11). Available via: http://dx.doi.org/ 10.1002/asi.20834 [accessed 5 April 2009].

Thelwall, M. (2009a, to appear). *Introduction to Webometrics.* New York: Morgan Claypool.

Thelwall, M., (M.Thelwall@wlv.ac.uk), (2009b). *RE:Query regarding LexiURL searcher.* E-mail to P. Wells (Paul.Wells@Anglia.ac.uk). Sent 6th April 2009.

Thelwall, M. and Kousha, K. (2008) Online presentations as a source of scientific impact? An analysis of PowerPoint files citing academic journals. *Journal of the American Society for Information Science and Technology* [online] 59(5). Available via: http://dx.doi.org/10.1002/asi.20803 [accessed 5 April 2009].

Thelwall, M., Vaughan, L. and Björneborn, L. (2005) Webometrics. *Annual Review of Information Science and Technology* [online] 39(1). Available via: http://dx.doi.org/10.1002/aris.1440390110 [accessed 28 December 2008].

Thomas, C. and McDonald, R. (2007) Measuring and Comparing Participation Patterns In Digital Repositories: Repositories by the Numbers, Part 1. *D-Lib Magazine* [online] 13 (9/10). Available via: http://dx.doi.org/10.1045/september2007-mcdonald [accessed 19 April 2009].

Thomas, C. and McDonald, R. (2008) In Search Of A Standardized Model for Institutional Repository Assessment or How Can We Compare Institutional Repositories? *Proceedings of the ARL 2008 Assessment Conference* [online] 3(10). Available via: http://repositories.cdlib.org/postprints/3240 [accessed 25 January 2009].

University of Bath (2008) *Institutional Repository Software candidates* [online]. Available via: https://wiki.bath.ac.uk/display/bucswebdev/Institutional+Repository+Software+candidates [Accessed 20 November 2008].

University of Nottingham (2008a) *The Directory of Open Access Repositories - Open*DOAR [online]. Available via: http://www.opendoar.org/index.html [accessed 7 February 2009].

University of Nottingham (2008b) *About Open*DOAR [online]. Available via: http://www.opendoar.org/about.html [accessed 7 February 2009].

University of the West of England (2007) *Research Observatory* [online]. Available via: http://ro.uwe.ac.uk/ [accessed 3 August 2008].

Vaughan, L. and Hysen, K. (2002) Relationships between links to journal Web sites and impact factors. *Aslib Proceedings* [online] 54(6). Available via: http://dx.doi.org/10.1108/00012530210452555 [accessed 7 March 2008].

Vaughan, L. and Thelwall, M. (2002) Scholarly use of the Web: What are the key inducers of links to journal Web sites? *Journal of the American Society for Information Science and Technology* [online] 54(1). Available via: http://dx.doi.org/10.1002/asi.10184 [accessed 21 November 2008].

Vaughan, L., Kipp, M. and Gao, Y. (2006) Why are hyperlinks to business Websites created? A content analysis. *Scientometrics* [online] 67(2). Available via: http://dx.doi.org/10.1007/s11192-006-0100-6 [accessed 5 April 2009].

Vaughan, L., Kipp, M. and Gao, Y. (2007) Are co-linked business websites really related? A link classification study. *Online Information Review* [online] 31(4). Available via: http://dx.doi.org/10.1108/14684520710780403 [accessed 20 March 2009].

VOSON (2008) *Welcome to the Virtual Observatory for the Study of Online Networks* [online] Available via: http://voson.anu.edu.au/ [accessed 28 January 2009].

Ware, M. (2004) Institutional repositories and scholarly publishing. *Learned Publishing* [online] 17(2). Available via: http://dx.doi.org/10.1087/095315104322958490 [accessed 17 January 2009].

Weber, R. (1990) *Basic Content Analysis.* 2nd ed. Newbury Park, California: Sage.

Westell, M. (2006). Institutional repositories: Proposed indicators of success. *Library Hi Tech* [online] 24(2). Available via: http://dx.doi.org/10.1108/07378830610669583 [accessed 28 September 2008].

Wheatley, P. (2004) Institutional Repositories in the Context of Digital Preservation. *Microform & Imaging Review* [online] 33(3). Available via: http://dx.doi.org/10.1515/MFIR.2004.135 [accessed 19 April 2009]

Wilkinson, D., Harries, G., Thelwall, M. and Price, L. (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* [online] 29(1). Available via: http://dx.doi.org/10.1177/016555150302900105 [accessed 21 February 2009].

Xia, Jingfeng (2008) A Comparison of Subject and Institutional Repositories in Self-archiving Practices. *Journal of Academic Librarianship* [online] 34(6). Available via: http://dx.doi.org/10.1016/j.acalib.2008.09.016 [accessed 2 February 2009].

Zuccala, A. and Thelwall, M. (2005) *Web Intelligence Report.* University of Wolverhampton: School of Computing and Information Technology.

Zuccala, A., Thelwall, M., Oppenheim, C. and Dhiensa, R. (2006) *Digital Repository Management Practices, User Needs and Potential Users: An Integrated Analysis* [online]. Available via: http://ie-repository.jisc.ac.uk/139/ [accessed 21 December 2008].

Zuccala, A., Thelwall, M., Oppenheim, C. and Dhiensa, R. (2007) Web Intelligence Analyses of Digital Libraries: A Case Study of the National electronic Library for Health (NeLH). *Journal of Documentation* [online] 63(4). Available via: http://dx.doi.org/10.1108/00220410710759011 [accessed 20 November 2008].

Zuccala, A., Oppenheim, C. and Dhiensa, R. (2008). Managing and evaluating digital repositories. *Information Research* [online] 13(1). Available via: http://InformationR.net/ir/13-1/paper333.html [accessed 28 October 2008].

7. Appendices

7.1 Appendix A

Example of excerpts from LexiURL searcher link impact report for Institution A.

Search Engine Results Report

Introduction

This report presents the results of a series of search engine queries, obtained from the first [50/1000/all] results returned by the search engine.

- Data from search engine: Yahoo! via its Applications Programming Interface
- Data gathered on: 07 January 2009.

Note that any queries with zero results are not shown anywhere in this report.

Overview of results

EstHits	Search	HitsReturned
5640	linkdomain:eprints.[redacted].ac.uk -site:eprints.[redacted].ac.uk	983

Detailed Results

- Results from base query: linkdomain:eprints.soton.ac.uk -site:eprints.soton.ac.uk URLs - Domains - Sites - Sites(domains) - STLDs - TLDs - Random domains

TLD summary

STLD summary

Home page

URLs of pages matching the base query: linkdomain:eprints.[redacted].ac.uk -site:eprints.[redacted].ac.uk

1. http://www.123people.de/s/kai+forster
2. http://www.dy-soton.150m.com/
3. http://www.dy-soton.150m.com/about_me.html
4. http://blog.163.com/capaa/
5. http://blog.163.com/dingaiwu139/
6. http://202.115.193.225/Article/readers/20070628145722.html
7. http://202.195.195.137:81/library/plus/view.php?aid=103
8. http://72.15.199.215/blogs/united_kingdom/default.aspx
9. http://www.gabyyhackerteam.3xforum.ro/
10. http://144.16.65.194/hpg/envis/doc99html/infognuep250101.html
11. http://www.8ung.at/omega_news/on_global_climate_disruption.htm
12. http://members.a1.net/grubigstein/VG/litlist_CLIL_links.htm
13. http://www.inf.aber.ac.uk/elecinfo/openaccess.asp
14. http://www.abovetopsecret.com/forum/thread418796/pg22
15. http://www.openu.ac.il/research_center/dagim/dagim6/
16. http://soton.academia.edu/KeithJones/Papers
17. http://accesstower.de/sozialer-wandel.html
18. http://accesstower.de/technisch-untersuchung.html
19. http://acscinf.org/docs/meetings/229nm/229cinfabstracts.htm
20. http://www.acted.co.uk/forums/showthread.php?p=7450
21. http://www.addthesite.com/HEALTH/Healthcare_Industry/
22. http://digital.library.adelaide.edu.au/dspace/handle/2440/40009

Home page

Top-level domains of pages matching the base query: linkdomain:eprints.[redacted].ac.uk -site:eprints.[redacted].ac.uk

The table lists the top-level domains of pages matching the base query: linkdomain:eprints.soton.ac.uk -site:eprints.soton.ac.uk. The Domains column lists the number of Domains returned by the query with the given STLD. Domains are equated with domain name segments, normally the part of the URL after the http:// and before the first subequent slash (e.g., www.bbc.co.uk).

TLD	Domains	%
com	227	29.5%
uk	132	17.1%
org	117	15.2%
net	45	5.8%
de	25	3.2%
edu	24	3.1%
cn	20	2.6%
info	18	2.3%
ca	14	1.8%
jp	13	1.7%
pl	13	1.7%
au	11	1.4%
fr	8	1.0%
br	6	0.8%

Home page

Second or top-level domains of pages matching the base query: linkdomain:eprints.[redacted].ac.uk -site:eprints.[redacted].ac.uk

The table lists the second or top-level domains of pages matching the base query: linkdomain:eprints.soton.ac.uk -site:eprints.soton.ac.uk. The Domains column lists the number of Domains returned by the query with the given STLD. Domains are equated with domain name segments, normally the part of the URL after the http:// and before the first subequent slash (e.g., www.bbc.co.uk).

STLD	Domains	%
com	227	29.5%
org	117	15.2%
ac.uk	97	12.6%
net	45	5.8%
co.uk	25	3.2%
de	25	3.2%
edu	24	3.1%
info	18	2.3%
ca	14	1.8%
pl	13	1.7%
edu.cn	12	1.6%
fr	8	1.0%
edu.au	8	1.0%
ac.jp	8	1.0%

Home page

Random domains from pages matching the base query: linkdomain:eprints.[redacted].ac.uk -site:eprints.[redacted].ac.uk

The table lists up to 400 random domains of pages matching the base query: linkdomain:eprints.soton.ac.uk -site:eprints.soton.ac.uk, together with a random URL from each domain.

No.	Domain	URL
1	www.psychology.soton.ac.uk	http://www.psychology.soton.ac.uk/people/showpublications2.php?username=timw
2	inundated.tescilc4.info	http://inundated.tescilc4.info/genet/hopeful.php
3	crofsblogs.typepad.com	http://crofsblogs.typepad.com/climate/2007/08/sun-is-not-a-fa.html
4	www.johnsonhomeplus.com	http://www.johnsonhomeplus.com/cooking/southernlivingcookbook/
5	www.wmo.ch	http://www.wmo.ch/pages/prog/wcrp/NewsArchives_index.html
6	www.noc.soton.ac.uk	http://www.noc.soton.ac.uk/nocs/research.php
7	szgy.org	http://szgy.org/user/ento/engine
8	uk.encarta.msn.com	http://uk.encarta.msn.com/TVEI.html
9	www.soudoc.com	http://www.soudoc.com/html/71/t-8676171.html
10	busca.igbusca.com.br	http://busca.igbusca.com.br/app/busca.ig?o=ULTIMOSEGUNDO&start=p&t=4&q=fan%E7a&d-4025704-p=4&gpage=4
11	penfold.lib.hull.ac.uk	http://penfold.lib.hull.ac.uk:8080/confluence/display/golddust/RSS+feeds
12	ljk.imag.fr	http://ljk.imag.fr/membres/Arthur.Vidard
13	network.nature.com	http://network.nature.com/forums/harvardpublishingforum/1047
14	my.dek-d.com	http://my.dek-d.com/PasZo/diary/?day=2008-06-14
15	www.freelists.org	http://www.freelists.org/archives/18650/08-2004/msg00003.html

7.2 Appendix B

Summary of content analysis for Institution A.

Address of web page containing in-link	Source page	Target page	Link motivation	Possible user group
http://www.psychology.[redacted].ac.uk/people/showpublications2.php?username=timw	Staff Publications Page - Same Institution	Item and Item Page	Ownership	Academic-same institution
http://www.johnsonhomeplus.com/cooking/southernlivingcookbook/	Automatically generated page/link spam	Item Page	n/a	n/a
http://www.wmo.ch/pages/prog/wcrp/NewsArchives_index.html	Departmental/ research group page - Other/ cross Institution	Item	Recommendation	Academic-different institution
http://www.noc.[redacted].ac.uk/nocs/research.php	Departmental/ Research Group Page - Same Institution	Technical Page	Ownership	Academic-same institution
http://penfold.lib.hull.ac.uk:8080/confluence/display/golddust/RSS+feeds	Wiki	Technical Page	Reference	Academic-different institution
http://ljk.imag.fr/membres/Arthur.Vidard/	Staff Publications Page - Different Institution	Item	Ownership	Academic-different institution
http://network.nature.com/groups/harvardpublishingforum/forum/topics/1047	Blog	Homepage	Example	Academic-different institution/ Repository related
http://liinwww.ira.uka.de/bibliography/Misc/eprints.[redacted].ac.uk.html#about	Library/ Repository related website - Other institution	Homepage	Ownership	Academic-different institution
http://iscte.pt/~jmgd/research.html	Staff Publications Page - Different Institution	Item Page	Ownership	Academic-different institution

http://www.abovetopsecret.com/forum/thread418796/pg22	Forum/ Discussion Board	Item Page	Reference	Public
http://www.papimi.gr/inflcan.htm	Automatically generated page/link spam	Item Page	n/a	n/a
https://mx2.arl.org/Lists/SPARC-OAForum/Message/1382.html	Email List Archive	Homepage	Reference	Academic-different institution/ Repository related
http://www.ljmu.ac.uk/lea/77337.htm	Library/ Repository related website - Other institution	Homepage	Recommendation	Academic-support/ Repository related
http://www.isvr.[redacted].ac.uk/Staff/staff18.htm	Staff Publications Page - Same Institution	Technical Page	Ownership	Academic-same institution
http://www.faunaclassifieds.com/forums/printthread.php?t=85549	Forum/ Discussion Board	Item Page	Reference	Public
http://americanschool.chambordmusic.com/italian_violin/index.htm	Automatically generated page/link spam	Item Page	uncertain	non-academic
http://www.addthesite.com/HEALTH/Healthcare_Industry/	Automatically generated page/link spam	homepage	n/a	non-academic
http://www.tuphoto.ru/displayimage.php/?drug-info=431	Automatically generated page/link spam	Item Page	n/a	non-academic?
http://www.benningtonenergy.org/article.php?article=20051018_001	Commercial/ Industrial-related webpage	Item Page	Reference	Professional
http://www.nichewatch.com/hydrodynamic_cavitation.html	Automatically generated page/link spam	Item Page	n/a	non-academic
http://vre.upei.ca/riri/node/64	Library/ Repository related website - Other institution	Item Page	Affiliation	Repository related

http://sjcssc.com/gointernational/displayimage.php/?pharm-drug=470	Automatically generated page/link spam	Item Page	n/a	non-academic? (automated)
http://www.nal.usda.gov/awic/pubs/Dogs/housing.shtml	Government-related webpage	Item Page	Reference	Professional
http://www.math.harvard.edu/~roed/writings/talks.html	Staff Publications Page - Different Institution	Item Page	Reference	Academic-different institution
http://www.sconul.ac.uk/groups/information_literacy/committee/ww.html	Library/ Repository related website - Other institution	Item Page	Ownership	Repository related
http://www.stats.gla.ac.uk/~paulj/publications.html	Staff Publications Page - Different Institution	Item Page	Ownership	Academic-different institution
http://www.nerc.ac.uk/about/access/statement.asp	Government-related webpage	Homepage	Affiliation	Professional
http://mustanggenerations.com/store/index.php/?rx-show=118	Automatically generated page/link spam	Item Page	uncertain	n/a
http://hackaday.com/2008/01/	Blog	Item Page	Reference	Public
http://www.utsc.utoronto.ca/~elpub2008/2008/06/opening-scholarship-workshop.html	Blog	Homepage	Example	Repository related
http://www.timkelf.com/PresentedWork.html	Staff Publications Page - Different Institution	Item Page	Ownership	Academic-different institution
http://www.horsetalk.co.nz/news/2008/07/056.shtml	Other Public webpage	Item Page	Reference	Public
http://listserver.sigmaxi.org/sc/wa.exe?A2=ind07&L=american-scientist-open-access-forum&D=1&F=I&O=D&P=63220	Email List Archive	Item Page	Reference	Academic-different institution

http://www.gfdl.noaa.gov/~gav/pubs.html	Staff Publications Page - Different Institution	item page	Ownership	Academic-different institution
http://www.openarchives.org/pipermail/oai-implementers/2005-February/001429.html	Email List Archive	Technical Page	Reference	Repository related
https://www.staffs.ac.uk/uniservices/student_office/new_student/facultyfirststepshomepage/AMDactivitypagefineart.php	Library/ Repository related website - Other institution	Item	Recommendation	Academic-support
http://healthguide.co.uk/library.html	Other Public webpage	Item Page	Recommendation	Public
http://www.propertyentrepreneur.me.uk/entrepreneur-stories	Automatically generated page/link spam	Item Page	n/a	n/a
http://eprints.kingston.ac.uk/information.html	Library/ Repository related website - Other institution	Homepage	Example	Repository related