Reflections on Three Years of Archiving Research Output

Hussein Suleman
University of Cape Town
hussein@cs.uct.ac.za

ABSTRACT

While “Open Access” and “institutional repositories” have received some attention in South Africa in recent years, very few tertiary institutions and departments have made a commitment to opening up access to their research. The Department of Computer Science at UCT was one of the first departments in the country to adopt the notion of an institutional repository, albeit localized, because of a strong open access tradition within the discipline. This repository of published and working papers and reports has evolved over a period of three years to become a core part of the operations of the department. It is arguably a successful implementation of an institutional repository. As such, this paper presents an analysis of the development of the archive and its use, providing anecdotal and statistical evidence to demonstrate that such projects are indeed feasible and result in substantial visibility of research for South African institutions and their academic departments.

Keywords

Open access, institutional repository, archive, digital libraries, E-prints, EPrints, publications.

INTRODUCTION

Open Archives and Open Access

The Department of Computer Science (CS) at the University of Cape Town (UCT) established a Publication Archive (http://pubs.cs.uct.ac.za) in the beginning of 2003 to store research- and teaching-related outputs of the department for internal and external consumption. While this paper discusses and analyses the archive in relative isolation, the establishment and subsequent success of the project is largely due to the context and conducive environment in which the archive was created. A number of converging factors have led to this conducive environment, including: the acceptance of the Open Archives Initiative Protocol for Metadata Harvesting (Lagoze, Van de Sompel, Warner and Nelson, 2001); the emergence of Open Access as a viable means for research dissemination; the popularity of Institutional Repositories; and the evolution of the Networked Computer Science Technical Reference Library (Anan, Liu, Maly, Nelson, Zubair, French, Fox and Shivakumar, 2002).

In the late 1990’s many large and international digital library/archive projects operated somewhat in isolation and there was a high degree of duplication of effort in creating such projects as their requirements were not aligned with one another. As the number of archives grew and needs of the scholarly community changed, it became apparent that there was a need to define common services, tools and standards. One of the first of such efforts was led by the Open Archives Initiative (OAI), which developed the Protocol for Metadata Harvesting (PMH) as a low-barrier to interoperability (Lagoze and Van de Sompel, 2001). This Web-based protocol defines how to transfer a stream of XML-encoded metadata from one machine to another such that data collections can be synchronised. The OAI-PMH rapidly became a key underlying technology for distributed digital libraries as well as scholarly search engines. The community of researchers and practitioners rallied around this standard and created a series of reusable tools, available from the OAI website, which may be plugged into existing or new digital library systems. More significantly, some institutions commissioned the building of complete Open Source software toolkits to manage digital collections of scholarly output for institutions or organisations. Prominent among these efforts was MIT’s DSpace system (Smith, Barton, Branschofsky, McClellan, Walker, Bass, Stuve and Tansley, 2003) and the University of Southampton’s EPrints system (University of Southampton, 2006a) – the latter was used for the UCT CS archive.

While interoperability and the availability of software tools for document management were important to archivists, many academics and institutions were at the same time feeling the effects of the “Serials Crisis”. Since the late 1970s, the prices for access to the journals that academics need in order to conduct their research have been increasing significantly faster than inflation (Suber, 2004). Institutions have had to cancel subscriptions to some of these journals because of limited funds and high costs. As these prices soared, disgruntled academics and librarians began to look at alternatives, and one such alternative presented itself at the launch of the Budapest Open Access Initiative (BOAI) (Open Society Institute, 2004), supported by the Open Society Institute (OSI). Here a number of leading academics from around the world gathered to declare that research literature should be made freely available on the Internet, giving impetus to and defining a growing Open Access movement.

Harnad, a leading advocate for Open Access (OA), has repeatedly characterised OA solutions as falling into either the "Green" or "Gold" road (Harnad, 2004). The Gold road is the ideal, where academics submit their research outputs only to journals which operate on an OA basis, i.e., free online access. Unfortunately, many budding scholars need to build their careers and this sometimes requires publishing in established traditional (closed) journals. The alternative, the Green road, involves the establishment of institutional repositories where academics can place copies of their peer-reviewed articles and papers as an alternative access route for readers. Harnad stresses that universities need to adopt this self-archiving approach because without access the products of their research will have no impact (Harnad, 2003). This approach requires that copyright is adhered to, but an increasing number of publishers allow the self-archiving of papers and articles in personal, department and university archives, either immediately after formal publication or shortly afterwards. The Sherpa project tracks the copyright policies of publishers and, as of 22 July 2006, 94% of journals assessed allowed self-archiving of either pre- and/or post-prints of articles (University of Southampton, 2006b).

A number of studies have demonstrated that OA leads to increased access to resources and an increase in the citation impact of the resources. The earlier Lawrence study (Lawrence, 2001) showed that there was a clear correlation between online access to articles in Computer Science and a higher number of citations. Lawrence further stresses that it is the availability of articles that increases the visibility of research and ultimately citation of these. Recently, Eysenbach (2006) studied articles that were OA compared to those that were not within a single journal, and concluded that OA led to a doubling of the citation rate in the first year and an even greater impact thereafter.

From a technological perspective, the OSI recommends that all software systems supporting OA also adhere to open standards, especially the OAI-PMH, and maintains a regularly-updated guide to institutional repository software to make it easier for institutions to choose among the various options available (Crow, 2004). Systems such as EPrints fit into this model and are indicative of the close relationship between the community of users supporting Open Access and those supporting open standards.

Motivation for a Publication Archive at UCT Computer Science

In this climate of Open Access, the UCT CS Publication Archive was set up primarily to archive technical reports electronically, with a secondary aim of hosting a permanent record of all publications emanating from the department.

Technical reports have historically been a means whereby computer scientists around the world have published detailed documents that would not naturally fit into the model of a research paper, but where the level of detail is required for future extensions or verification of the work. Many departments have published these documents online to allow access to researchers elsewhere.

The Networked Computer Science Technical Reference Library (NCSTRL) is an international project that aims to collate metadata related to technical reports from different departments into a single central discovery system (Leiner, 1998). Visitors to the NCSTRL website are able to browse or search through the meta-collection of technical reports across all institutions and thereafter they are able to download the resources directly from their source archives. The success of the principle behind NCSTRL, but the limited success of its somewhat proprietary technology, contributed to the development of the OAI-PMH. As a prime driver, NCSTRL was then one of the first projects to adopt the OAI-PMH as the basis for setting up its distributed digital library environment. Thus, any institution operating an institutional repository (as per the OSI definition) can almost trivially participate in NCSTRL.

The UCT CS Publication Archive uses one of the recommended institutional repository software packages – EPrints – to enable both local archiving as well as the ability to participate in international communities such as NCSTRL and the equivalent for Electronic Theses and Dissertations, NDLTD (Suleman, Atkins, Goncalves, France, Fox, Chachra, Crowder and Young, 2001).

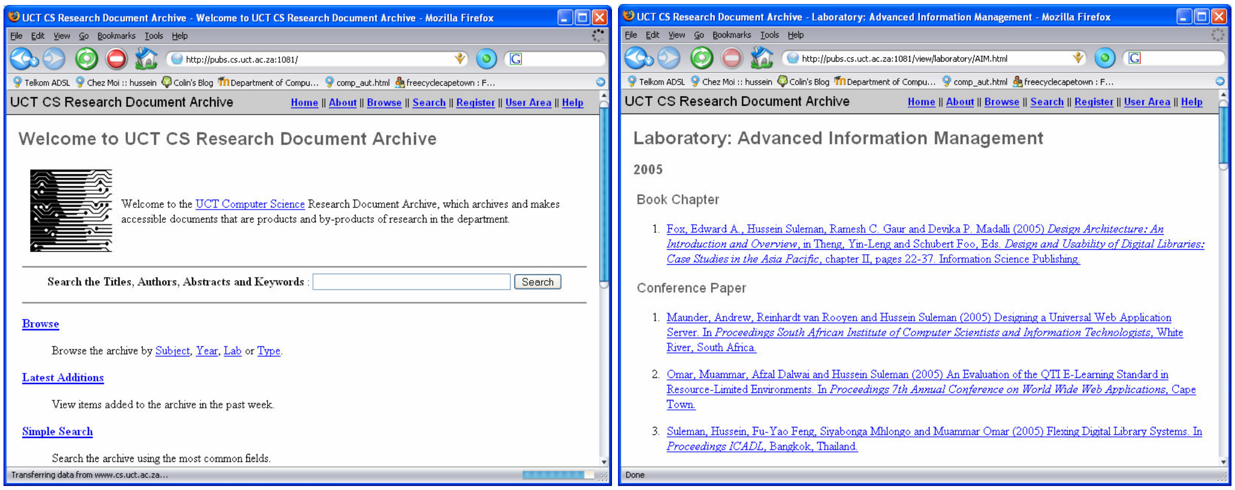
THE PUBLICATION ARCHIVE

Core EPrints Features

The following is a list of the core features provided by the EPrints software as utilised by the UCT CS archive:

- Documents are submitted by their authors or designated individuals, such as publications officers within some of the research laboratories. Each user is able to create and manage a user account, solely for the purpose of submission, without the intervention of an administrator.
- During submission, a user first enters metadata – which differs depending on what type of resource is being submitted – and then submits the resource in PDF format, and optionally other formats. Finally, the user needs to make a rights declaration and review the entire submission before submitting it for review by the editor.
- The editor carefully scrutinises all submissions for completeness and correctness of both the metadata and digital objects before accepted them into the archive. The editor of the archive confirms that the submissions are both suitable and relevant.
- Once a resource has been accepted into the archive, it can be found through the search interface, which allows users to search for documents by providing a simple term list or a fielded query.
- Documents are also classified into a subject hierarchy based on the top two levels of the ACM Computing Classification System (ACM, 1998). Documents may be browsed on the basis of subject, laboratory, year or type.
- The OAI-PMH can be used to harvest metadata on an incremental basis from the archive. This is a machine-to-machine interface for interoperability purposes.

Figure 1 shows the front page of the archive and a listing of entries for one of the research laboratories.


UCT CS Research Document Archive - Welcome to UCT CS Research Document Archive - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://pubs.cs.uct.ac.za:1081/
Telkom ADSL Chez Moi :: hussein Colin's Blog Department of Compu… comp_aut.html freecyclecapetown : F…
UCT CS Research Document Archive
Home || About || Browse || Search || Register || User Area || Help
Search
Transferring data from www.cs.uct.ac.za...

Welcome to UCT CS Research Document Archive

Welcome to the UCT Computer Science Research Document Archive, which archives and makes accessible documents that are products and by-products of research in the department.

Search the Titles, Authors, Abstracts and Keywords :

Browse

Browse the archive by Subject, Year, Lab or Type.

Latest Additions

View items added to the archive in the past week.

Simple Search

Search the archive using the most common fields.

UCT CS Research Document Archive - Laboratory: Advanced Information Management - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://pubs.cs.uct.ac.za:1081/view/laboratory/AIM.html
Telkom ADSL Chez Moi :: hussein Colin's Blog Department of Compu… comp_aut.html freecyclecapetown : F…
UCT CS Research Document Archive
Home || About || Browse || Search || Register || User Area || Help
Done

Laboratory: Advanced Information Management

2005

Book Chapter

1. Fox, Edward A., Hussein Suleman, Ramesh C. Gaur and Devika P. Madalli (2005) *Design Architecture: An Introduction and Overview*, in Theng, Yin-Leng and Schubert Foo, Eds. *Design and Usability of Digital Libraries: Case Studies in the Asia Pacific*, chapter II, pages 22-37. Information Science Publishing.

Conference Paper

1. Maunder, Andrew, Reinhardt van Rooyen and Hussein Suleman (2005) Designing a Universal Web Application Server. In *Proceedings South African Institute of Computer Scientists and Information Technologists*, White River, South Africa.
2. Omar, Muammar, Afzal Dalwai and Hussein Suleman (2005) An Evaluation of the QTI E-Learning Standard in Resource-Limited Environments. In *Proceedings 7th Annual Conference on World Wide Web Applications*, Cape Town.
3. Suleman, Hussein, Fu-Yao Feng, Siyabonga Mhlongo and Muammar Omar (2005) Flexing Digital Library Systems. In *Proceedings ICADL*, Bangkok, Thailand.

Figure 1. Front page and browsable list of citations in UCT-CS Publication Archive

Local Customisations

The UCT CS archive was initially set up by customising parameters of the EPrints software. This included the following:

- A field was introduced for the research laboratory – this makes it possible to view the documents produced within a particular research grouping.
- The OAI-PMH interface was programmed to support both the RFC1807 metadata format (for the NCSTRL community) and the ETD-MS format (for the NDLTD community).
- The metadata elements were chosen to support a superset of allowable digital object types, including, for example, theses and conference papers.

In addition to the allowable customisations, the EPrints software was modified to better meet the needs of the UCT-CS department.

- Document listings are formatted to most closely resemble citation lists, with headings for years and subheadings for type of document – the version of EPrints available in 2003 did not allow for hierarchical sets of headings within a list of citations.
- The citations were themselves reformatted to better resemble a standard reference format – the default representation in EPrints does not match any popular format.

Roles of the Archive

The archive serves different roles within the department, including as:

- A local copy of all research output to which students and colleagues can be referred.
- A list of publications for each research laboratory.
- A list of research output of the entire department, for annual reporting to funding agencies and other authorities.

POPULATING THE ARCHIVE

Obtaining the very first submissions was a difficult task because students and staff had to be convinced of the utility of such an archive. Staff members were first approached to submit their documents but there was some hesitance in spite of the popularity of online technical report archives at Computer Science departments elsewhere. Some changes were made to the EPrints software to address possible concerns of staff. Once all changes were made, staff members were still reluctant to submit their documents because it was a change in their routines.

The most successful strategy to encourage participation was to get some documents into the archive and have it officially sanctioned by the department (linked off the departmental website). Immediately, the reputations of some staff were being promoted by virtue of their documents being present in the official archive, while others' reputations were being negatively affected by implication. Staff in research laboratories eventually nominated students in each of the research laboratories to coordinate the process. This approach has worked very well – although a large number of submissions do still come directly from authors.

The very first electronic thesis submitted was a milestone because students were wary of publishing their precious intellectual property online and possibly jeopardising their ability to publish in the traditional journals and conferences avenue and/or apply for patents. In the thesis and dissertation community, these fears are known to be unfounded because publishers do not consider non-peer-reviewed electronic theses to be in the same category as other forms of scholarly publication. Once the first document was eventually submitted and archived, others followed in domino-style, largely because no student would want to be the one whose research was not listed.

Every submission to the archive has been routinely screened by an editor. A large number of submissions initially contained incomplete metadata and these were returned to the submitters with comments on what to fix or add – in the 3rd year of the archive, this is hardly necessary any longer. Similarly, in the initial stages some editing was necessary to fix and standardise the representation of some metadata fields. A handful of submissions were rejected because copyright clearance was not obtained.

3 years later, most research groupings now refer to the archive instead of maintaining separate listings on their websites. Thus, submission and referrals to the archive have become a standard part of the research routine in the department. The archive still fulfils its primary purpose of a persistent electronic store for technical reports. Each year a major percentage of submissions to the archive originate in Honours project, where a technical report is a requirement of the department.

Figure 2 shows the number of submissions for each month of a 36 month period. The highest number of submissions corresponds to the Honours student technical reports, which are submitted around October of each year. In general, there is then a relatively consistent but small number of submissions of what are largely papers and articles each month. Some of the earlier months do not follow this trend because the different research groupings populated the archive with recent publications in batches, but not all at the same time.

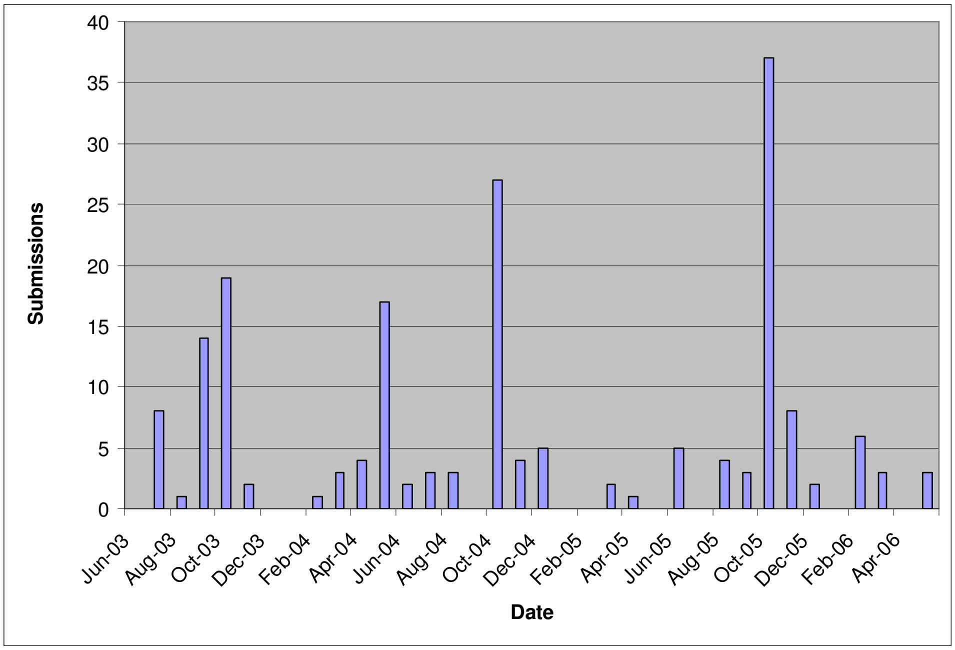


Figure 2. Number of submissions to the archive over time

ARCHIVE ACCESS

The archive has a complete set of server log files and metadata associated with each resource. This usage data has given rise to various forms of analysis, which confirm usage expectations of such an archive and reveal some interesting patterns and phenomenon. The analysis was limited to access patterns since others have already demonstrated that access is the most important criterion to increase impact. Also, these statistics have attempted to make as few assumptions as possible, with the caveat that any analysis of log files is an approximation at best.

In total, the archive contains 187 publicly-accessible resources (as of 22 July 2006), each of which consists of a reasonably complete metadata record and one or more attached files. A very small number of the submissions contains no attached file – because of copyright reasons. The remaining resources contain at least a PDF version of the resource. PDF was chosen as a standard format to ensure a degree of preservation and continued access over time, independent of the underlying platform.

The time period for the following analyses was June 2003 to May 2006, inclusive. This amounted to a raw number of 478520 accesses to files through the Web server, and logged in the server log file.

User Agents, Bots and Crawlers

The first step in characterising user behaviour based on log files is usually to separate the log file into those entries that were the result of human activity and those that were the result of crawlers and harvesters. Some machine-based agents, such as download managers, are considered to be mediated human activity.

User agent information was extracted from the log file and a unique list generated. User agents are typically unique names given to the HTTP client programs that make requests to the Web servers on behalf of users. Table 1shows the top 10 entries from this list. A large number of entries were tagged as being a flavour of Mozilla – without the Mozilla entries, there were a total of 798 unique user agents. The user agents were separated into those that were obviously crawlers and harvesters and those that were probably not. Obviously this distinction is not always clear as an HTTP client could easily masquerade as Firefox or IE. In this vein, some of the Mozilla entries were actually listed as "compatible" but in fact originated from crawlers – 2 entries in the table show this effect. A list of acceptable agents was created by manually removing crawler/harvester entries for all user agents with more than a non-trivial number (10) of entries in the log file.

66339 OAIHarvester/2.0 (www.sigla.ru; alex@lib.msu.ru)
60170 Mozilla/5.0 (compatible; Yahoo! Slurp; http://help.yahoo.com/help/us/ysearch/slurp)
46392 Googlebot/2.1 (+http://www.google.com/bot.html)
15469 Mozilla/5.0 (compatible; Googlebot/2.1; +http://www.google.com/bot.html)
13538 msnbot/1.0 (+http://search.msn.com/msnbot.htm)
13102 Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1)
11300 Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1; .NET CLR 1.1.4322)
10214 Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1)
8769 Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; .NET CLR 1.1.4322)
8169 Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.0)

Table 1. Most popular user agents encountered, and occurrence counts for each

After separating the log file, 65% of the total entries were attributable to crawlers and harvesters. This is a significant percentage and indicates the high degree to which the archive is visible to and indexed by search engines. Besides the traditional Web search engines, 34 of the user agents were some form of OAI harvester – this indicates that the archive is in fact part being accessed through this mechanism as well as through the static Web pages generated by EPrints.

Source of Accesses

Figure 3 shows the number of unique IP addresses that accessed resources in the archive in each month of the 3 year period, excluding crawlers and harvesters. These are broken down into accesses from within UCT (Local Access) and accesses from anywhere in the rest of the world. In each case, the number of PDF document hits is counted, in addition to a total number of hits to all types of files (including both HTML files and PDF files). Multiple hits from a single IP address within a single month is assumed to a single user – this is a conservative estimate, given that proxy servers will register as a single access point for many users.

Local access is somewhat erratic, with quieter periods in April, July, September and December. This correlates with vacation periods at the university, with the exception that September is not always quiet. Thus deviation is most probably due to Honours projects being completed soon after the September vacation.

Non-local access is more uniform and the level of access is significantly higher. The absolute numbers of visitors from outside the institution each month is a good indicator for the increased visibility of the department because of its archive. The trend of increasing visitors each month may also relate to the size of the archive in the time periods indicated.

In terms of PDF file access and HTML file access, the non-local users exhibit a similar pattern for each category. This means that the number of Web pages browsed per PDF file downloaded is approximately a constant ratio. Their combination of search, browse and document retrieval activities are therefore more consistent over time. The local users appear to access more HTML documents at times. This is possibly due to their changing requirements over the year – for example, Honours students could be looking for a list of citations in October when they submit their projects to ensure that they are listed. The small number of accesses do, however, render most analyses of this data statistically insignificant.

Figure 4 shows the number of unique IP addresses that originated from crawlers and harvesters, therefore are indicators of search engine activity. The simple trend shows that an increasing number of search engines are indexing the archive. There were spikes at some points in time, probably indicating the emergence of new services, a change in the crawling/harvesting schedules of existing services, or a change in the machines or number of machines performing the harvesting/crawling operations. The closely-related curves for PDF files and All hits indicates that search engines are consistently obtaining both the HTML and document files – with about 1/3 of all accesses being to PDF files.

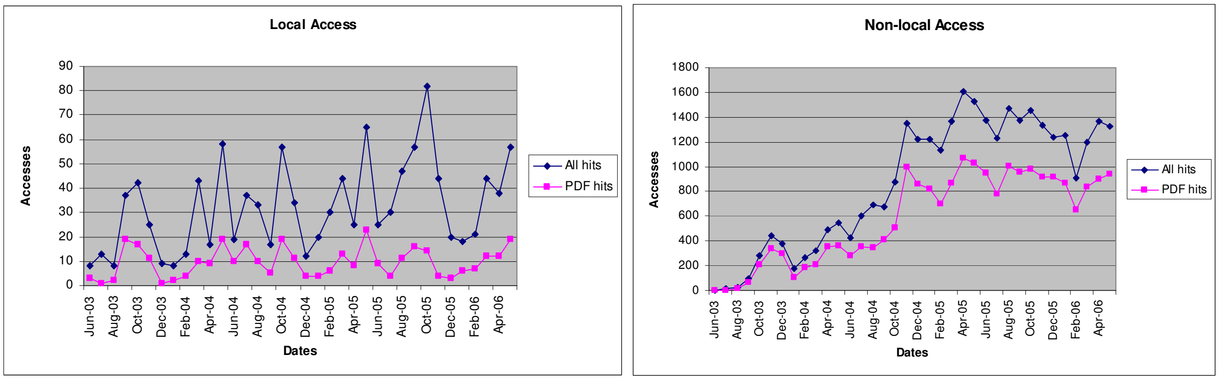


Figure 3. Number of unique user IP addresses connecting to the archive

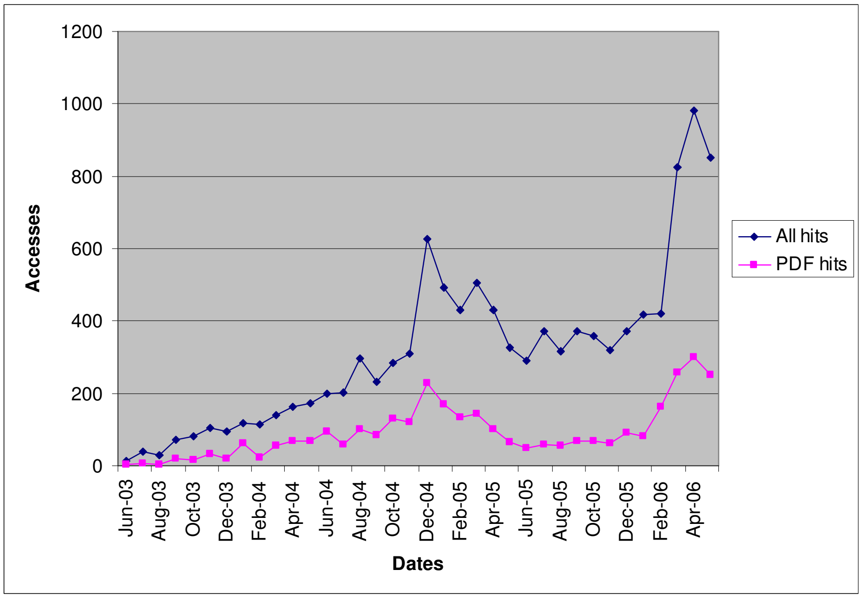


Figure 4. Number of unique IP addresses of crawlers and harvesters connecting to the archive

Arrival at Archive

Typically, digital libraries and institutional repositories provide elaborate search and browse systems but it is not clear that these are useful when the resources can be found through search engines. In an attempt to better understand the route users use to get to resources, the Referer field of the log file was used to classify where the user was coming from before accessing a particular PDF version of a resource.

Figure 5 illustrates the different routes users have taken to get to resources – the resources are ordered from least accessed to most accessed for clarity.

Google accounts for a large percentage of the source of document access, arguably a higher percentage for more popular documents. Other search engines show a similar trend to Google but with lower percentages of access

Pubs, which refers to users accessing a document after first browsing through the archive, has a more even effect on document access, implying that the population of users is more or less constant in size. These could be the same local users who are well acquainted with the front page of the archive or the departmental website which links to it.

Direct access to documents is because of no source information – these show a trend that is similar to the search engines and probably indicates that those Web browsers did not submit referer information and/or direct links were placed on Web pages or other digital library systems.

Popular documents have a very high percentage of referer entries from Google. For example, resource #199 is accessed 634 of 994 times by Google, thus it gets hit twice as many times because of Google than all other factors combined. The other factors are, however, all significant (Pubs=66, Direct=281, Other=13). This holds true for other popular documents as well.

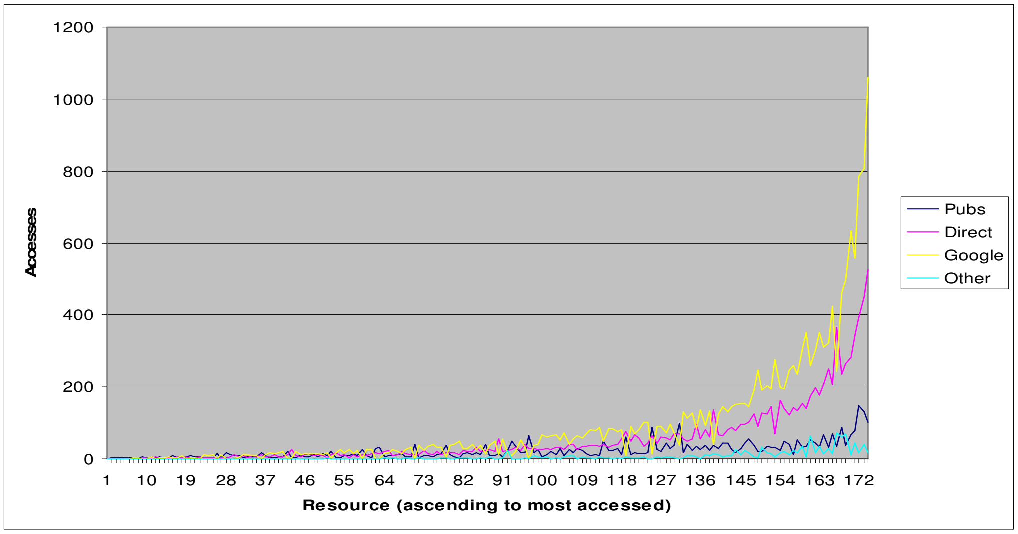


Figure 5. Access to resources by source/route

Access to Resources

Figure 6 shows the number of unique resources that have been accessed in each month, either from local users, non-local users or robots (crawlers/harvesters). These are compared directly with the total number of resources that were present in the archive in each month.

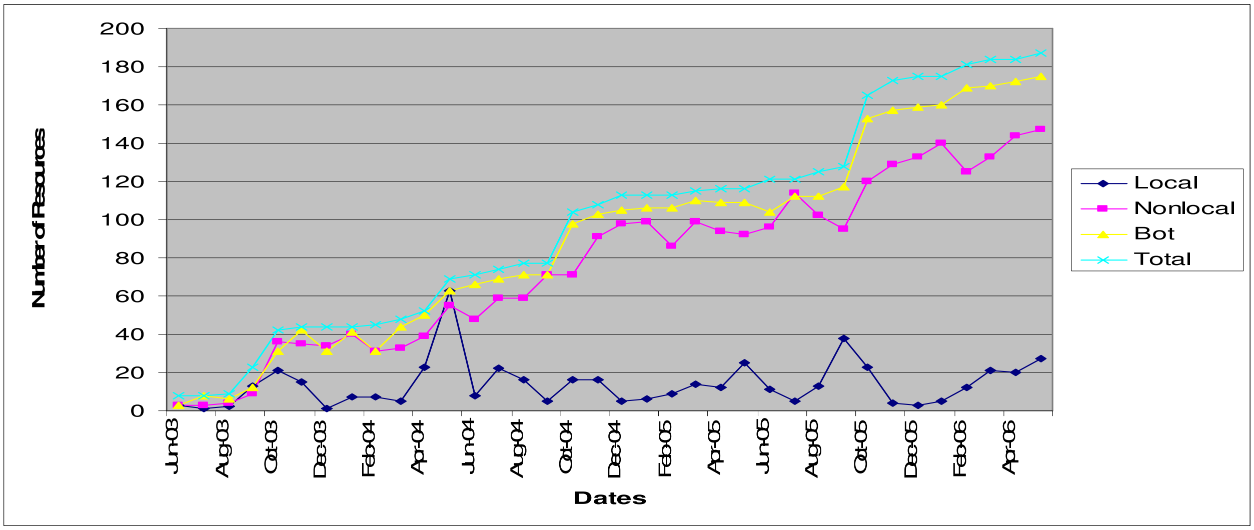


Figure 6. Number of unique resource accesses in each month

The local access shows a peak in the middle of each year, consistent with the academic calendar. Non-local users access the majority of resources in each month – almost every resource is viewed by some non-local users in every month. Robots are additionally even more aggressive in accessing close to the maximum number of available resources in every month. Thus, new resources are almost immediately being picked up by search engines, and this probably accounts for the equally high levels of access from non-local users.

Per Resource Access

A more thorough analysis could consider the use of resources on an individual basis. For this analysis, separating use of resources into local and non-local classes showed that local users never use any resource consistently over time. In almost all cases, the average number of uses per month of any single resource is zero (the average was only non-zero for resources added in the last month where the number of accesses in that first month was equal to the average). Some of the most highly requested resources (overall) were #149, #179 and #199. #149 was used only 8 times over its 17 months, #179 was used only 2 times over its 17 months and #199 was used only 7 times over its 22 months.

Thus, because of the relatively small contribution of local access, analysis of usage on an individual basis was conducted without separating the accesses into these classes. Table 2 shows the list of most frequently accessed resources, along with their titles and the average number of accesses per month (from unique users). To explain the popularity of these resources, it is clear from the titles that a few are surveys of their areas. Some of the other papers are popular because information on the topics they cover is highly sought at the time of writing.

#	Average	Title
149	68	Quality of Service of IEEE 802.11e
179	52	Online Course Material Interoperability and Tutorial Module for Moodle
199	52	Model Driven Communication Protocol Engineering and Simulation based Performance Analysis using UML 2.0
131	45	A Tutorial on RAID Storage Systems
139	25	Digital Rights Management - An Overview of current challenges and solutions
53	24	ChattaBox: A Case Study in Using UML and SDL for Engineering Concurrent Communicating Software Systems
166	23	Teaching Linux based Operating System
51	21	Using Voice over IP to Bridge the Digital Divide - A Critical Action Research Approach
78	21	Low-Cost Virtual Reality System (PS2-driven)

Table 2. Most popular resources in the archive

If the total numbers of downloads is considered instead of the per-month average across all months, the listing of resources is similar. The most downloaded resource is #149, at 1370, followed by #131 at 1128 and #179 at 1058 downloads respectively.

Figure 7 shows the access patterns for a sample of the items, those that are most popular in the archive and those that are oldest. The most popular items have a burst of interest, then consistently high levels of access. The oldest items have a more settled pattern but interest in older items does not diminish significantly over time!

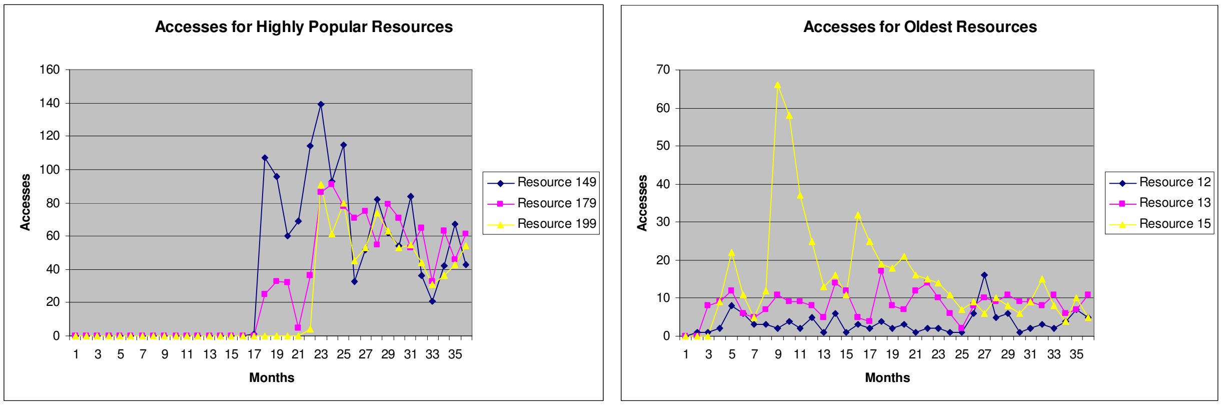


Figure 7. Accesses for most popular and oldest resources in each month

Figure 8 shows the popularity of resources based on the sequence of submission. The resources to the left were submitted before the ones to the right. The figure illustrates that the sequence of submission has no perceivable impact on the popularity. Resources submitted early on are not favoured and neither are newer resources. The ones that are downloaded can be assumed to be popular because of their presumed content.

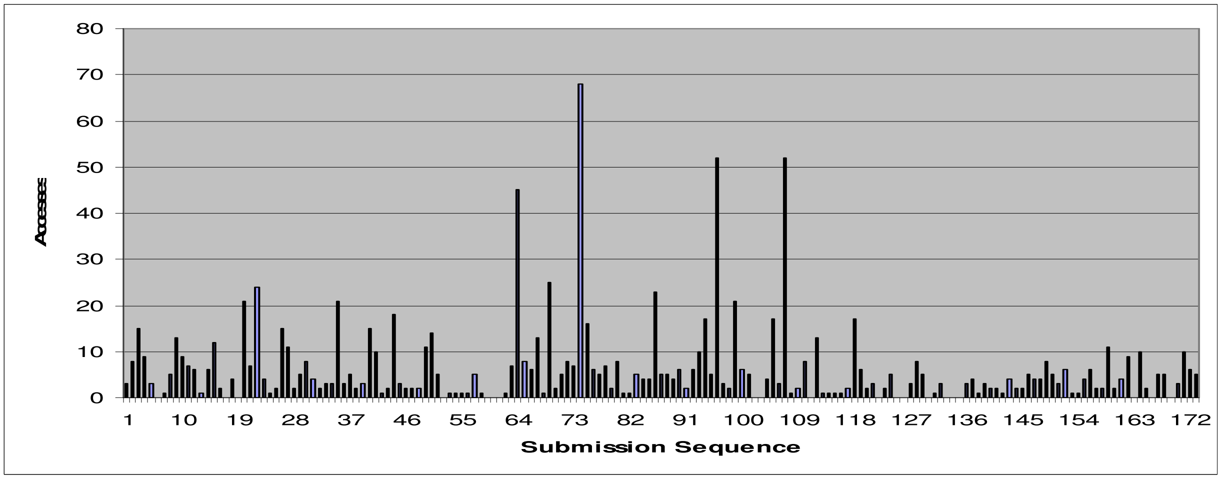


Figure 8. Popularity of resources based on their submission sequence

Figure 9 shows the distribution of resources by accesses per month with a logarithmic scale for the size of each access category. Most resources are accessed at least a few times per month, with a global average of 6.24.

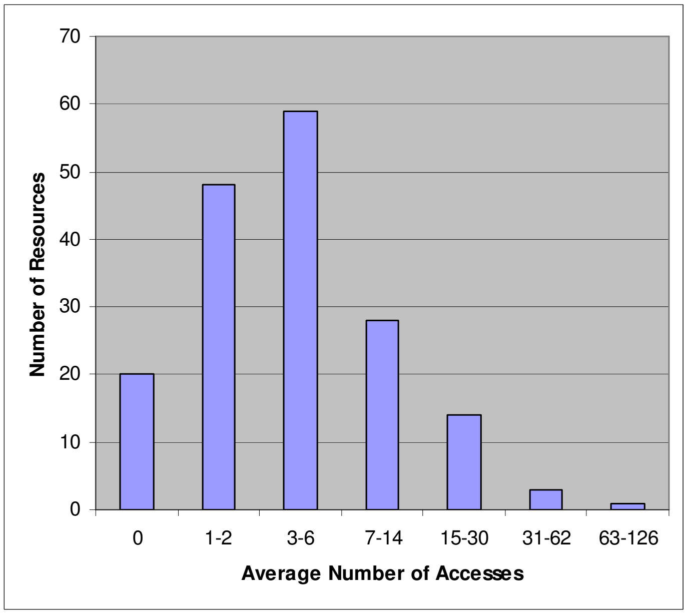


Figure 9. Distribution of accesses to resources

Downloading Experience

Finally, after finding a resource, users may or may not successfully download the resource. Modern browsers, in many instances, transparently handle continuation of downloading operations so the user does not need to bother with retrying a failed download and so that partial downloads can be concatenated to produce the complete resource as originally requested.

It may be contended that researchers outside South Africa cannot download resources from local institutional repositories because of bandwidth problems. In order to assess this, the HTTP status codes were extracted from the log and characterised in terms of being either a complete initial download or a partial download (either the first part or a subsequent part). Figure 10 shows the probability of obtaining a complete file on the first attempt, as well as the % of partial responses generated in comparison to the file sizes of resources.

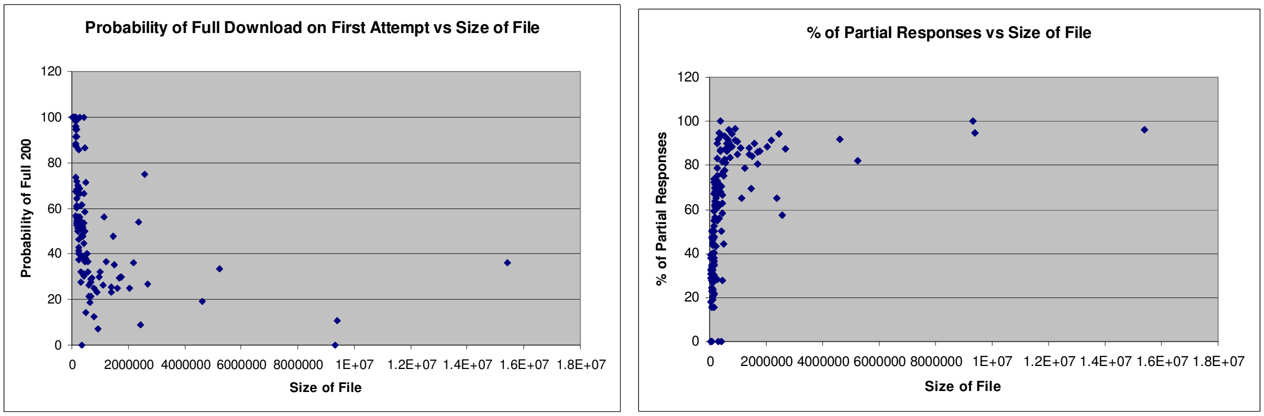


Figure 10. Success and partial response rates based on file size

The first analysis indicates that as the size of the file increases the probability of success on the first request decreases. Failure will result either in the user going away or the browser attempting a continuation of the transfer. The second analysis indicates that the number of partial responses increases with an increase in file size. This confirms that users have to make more attempts to transfer large files, as they do not always succeed on the first attempt.

SUMMARY OF RESULTS

In summary:

- Populating the archive has been successful and the archive is a routine part of departmental activity.
- The archive is regularly and aggressively indexed by search services.
- The vast majority of access to resources comes from outside UCT.
- Most users find their way to resources through Google and other search engines.
- Most resources are used in most months – by both crawlers/harvesters and end users.
- Access appears to be dictated by content, and not age of resources.
- While many downloads are successful immediately, larger files result in more partial downloads.

CONCLUSIONS

The UCT-CS Document Archive has served as a demonstration of the possibilities inherent in an Open Access archive serving the research and teaching needs of a single department within a university. The statistics derived from its log files, while fairly simple, illustrate the importance of archiving to immediate substantive visibility of research in an international context, and high visibility through the popular mechanisms used by researchers for locating resources (including scholarly archives that are typically OAI-based and general-purpose search engines such as Google).

Operating an archive for publications has resulted in high visibility and access to the research outputs of the department. It is hoped that the UCT CS experiences and the analysis of effectiveness of access to the UCT CS archive will encourage others to establish similar projects in their departments, faculties and universities.

ACKNOWLEDGMENTS

Thanks are due to UCT, the NRF and the Telkom/Siemens/THRIP Centre of Excellence for funding that supported this research.

REFERENCES (300)

1. ACM. 1998. Computing Classification System, ACM. Available http://www.acm.org/class/1998/
2. Crow, R. 2004. A Guide to Institutional Repository Software v3.0, Open Society Institute. Available http://www.soros.org/openaccess/software/
3. Eysenbach, G. 2006. Citation Advantage of Open Access Articles, *PLoS Biology*, 4(5): e157. Available http://biology.plosjournals.org/perlserv/?request=get-document&doi=10%2E1371%2Fjournal%2Epbio%2E0040157
4. Harnas, S. 2003. Maximizing university research impact through self-archiving, *International Journal on Science Communication*, 7, December 2003. Available http://jekyll.comm.sissa.it/articoli/art07_01_eng.htm
5. Harnad, S. 2004. The Green Road to Open Access: A Leveraged Transition, *American Scientist Open Access Forum*, 7 January 2004. Available http://www.ecs.soton.ac.uk/~harnad/Temp/greenroad.html
6. Anan, H, Liu, X., Maly, K., Nelson, M. L., Zubair, M., French, J. C., Fox, E. A., and Shivakumar, P. 2002. Preservation and transition of NCSTRL using an OAI-based architecture, in *Proceedings of JCDL 2002*, 14-18 July, Portland, OR, USA, ACM Press, 181-182.
7. Lawrence, S. 2001. Free online availability substantially increases a paper's impact, *Nature*, 31 May 2001. Available http://www.nature.com/nature/debates/e-access/Articles/lawrence.html
8. Lagoze, C., and Van de Sompel, H. 2001. The open archives initiative: building a low-barrier interoperability framework, in *Proceedings of JCDL 2001*, 24-28 June, Roanoke, VA, USA, ACM Press, 54-62.
9. Lagoze, C., Van de Sompel, H., Nelson, M. and Warner, S. 2002. The Open Archives Initiative Protocol for Metadata Harvesting, Open Archives Initiative. Available http://www.openarchives.org/OAI/openarchivesprotocol.html
10. Leiner, B. 1998. The NCSTRL Approach to Open Architecture for the Confederated Digital Library, *D-Lib Magazine*, 4 (11), December 1998. Available http://www.dlib.org/dlib/december98/leiner/12leiner.html
11. Open Society Institute. 2002. Budapest Open Access Initiative. Available http://www.soros.org/openaccess/read.shtml

12. Smith, M., Barton, M. R., Branschofsky, M., McClellan, G., Walker, J. H., Bass, M. J., Stuve, D., and Tansley, R. 2003. DSpace: An Open Source Dynamic Digital Repository, *D-Lib Magazine*, 9(1). Available http://www.dlib.org/dlib/january03/smith/01smith.html
13. Suber, P. 2004. A Primer on Open Access to Science and Scholarship, *Against the Grain*, 16(3).
14. Suleman, H., Atkins, A., Goncalves, M. A., France, R. K., Fox, E. A., Chachra, V., Crowder, M., and Young, J. 2001. Networked Digital Library of Theses and Dissertations: Bridging the Gaps for Global Access - Part 1: Mission and Progress, *D-Lib Magazine*, 7(9). Available http://www.dlib.org/dlib/september01/suleman/09suleman-pt1.html
15. University of Southampton. 2006. EPrints Free Software. Website http://www.eprints.org/software/
16. University of Southampton. 2006. Journal Policies - Summary Statistics So Far. Website http://romeo.eprints.org/stats.php