Institutional Repositories in New Zealand: An Analysis of Coordinators' Perspectives

by

Gregory James Benseman

Submitted to the School of Information Management,
Victoria University of Wellington
in partial fulfilment of the requirements for the degree of
Master of Information Studies

October 2012

Contents

Word Count of Body: 12,279

Abstract.....	2
Introduction.....	2
Background to study.....	4
Literature Review.....	5
1. Depositor attitudes to Repositories.....	5
2. Harvesting and Discovery of Repository Documents.....	6
3. Linked Data.....	7
4. Copyright Issues.....	9
Purpose of study.....	10
Theoretical Design	11
Research Objectives.....	12
Research Questions.....	13
Research Design/ Methodology.....	13
Limitations.....	17
Analysis.....	19
- Research Question 1: Collection methods: Statistics, Promotion.....	20
- Research Question 2: Description methods: Marsden, ANZSRC, DC.....	24
- Research Question 3: Collaboration: LCONZ, DNZ, CAIRSS.....	28
- Research Question 4: RMS integration: Sherpa, PBRF, Symplectic.	31
Conclusion, Suggestions for further research.....	37
Glossary.....	42
End Notes.....	44
Bibliography.....	46
Appendices.....	50
A. Information Sheet.....	51
B. Consent Form.....	53
C. Sample transcript page.....	55
D. Summary of the 2/5/11 IR Coordinators Discussion	56
E. NZ Research Holdings Counts.....	59
F. NZ metadata dashboard averages and complete DC field scores by Institution..	60

Abstract

This study is a small scale qualitative survey of coordinators working in institutional repository development in New Zealand since critical mass was reached in 2009. It aims to summarise their opinions on the current and future roles of their repository as both a preservation archive, and a discovery resource representing their institution's research community. The research uses narrative development techniques within the interpretivist paradigm to provide a contextual analysis of the repository's relationship with other repositories and the National Library. It is supported by quantitative analysis of the sampled repositories' holdings and the metadata quality with which the holdings are endowed. The analysis finds that since the establishment of New Zealand repositories, coordinators have adapted their collection strategies to encourage depositors towards Open Access publishing. These findings are placed in the context of the growth of non-mandated repository holdings and the technical infrastructure for harvesting resources, and integrating workflows with university research management systems. The results are used to discuss the goals coordinators have for improving the efficiency and visibility of their repository.

Introduction

Over the last decade the Open Access (OA) movement has promoted Institutional Repositories (IRs) as a way for universities to publicly present material donated by their research communities, as full text digital objects. However, during the great proliferation of IRs from 2006 – 2011, two issues, predicted even before Open Access first gained momentum in 2004, have become the leading cause of researchers ignoring their IR as the first place of publication.

The first issue revolves around copyright. During the Serials Crisis of 2007 - 2009, shrinking budgets forced libraries to find new approaches to subscribing to the journals that their communities used. Open Access supporters argued that by switching some funding to support subject based initiatives such as the post-print repositories arXiv or CogPrints, libraries (and IRs) and journal publishers would come to a more equitable arrangement for where much of the cost for refereeing and editing work

is borne by the institutional consortiums. This arrangement, known as Gold Access, is expected to be an intermediate step towards Green Access, where post-print material is published free of charge in digital repositories.

While certain publishers such as Springer and Elsevier have significantly lowered fees for Green Open Access publishing in IRs in the last four years, many academic disciplines outside of mathematics and the sciences lack the representation of large United States or European consortia to produce effective change. This is reflected in the research staff of those areas, that either remain ignorant of their local IR's existence, or are unwilling to jeopardize their careers by donating post-print articles to their IR. In the Finch report of June 2012 the transition¹ to Open Access is discussed in terms of the incredibly tangled relationship between publishers and universities. The Finch group projects a 50% uptake of Gold Access in the UK, compared to a 25% uptake globally².

This transition implies a massive overhaul of the academic publishing infrastructure, whereby the consortiums representing the national and international research communities are expected to rationalise the cost of producing and providing access to their pools of publicly funded research. In this process, the institutional repository's role is to act as a digital preservation archive of all the material its coordinator can obtain.

The second issue lies in promoting the viability of IRs as a useful public store, given the material they actually contain. Some institutional repositories only upload full text documents using Open Access (OA) licences such as Creative Commons. This can encourage depositors to use OA licences, or it can discourage them depending on a number of factors discussed in the literature review. Another institutional policy can be to upload some items to a restricted archive. The last policy option is to store items with all rights reserved licences as fragments in a dark archive, only accessible by coordinators at the institution's repository, and research department. Dark archives contain a range of material such as abstracts, data sets, preprint journal articles or books, and grey literature such as news items and magazines.

Background to Study

Since the initial Tertiary Education Commission Funding grants in 2006, New Zealand's Institutional Repositories have grown to what is considered critical mass of over 1,000 records. The most common repository software system is DSpace which runs with the LAMP package of storage, retrieval, ingestion, and administration tools. The DSpace community includes more than 1,000 repositories around the world, most of which use version 1.7 which is designed for XMLUI (Manakin) customisation, allowing repository coordinators to configure their institution's search interface's sidebars and metadata displays.

Repository coordinators must be aware of the plethora of material available, and the copyright and resourcing obstacles to collecting and presenting it. The main problem is how to encourage researchers to deposit work as full text OA objects either with the research department, or through the repository interface.

In New Zealand, most IRs are maintained by between 0.5 and 4.0 FTE staff. For many coordinators this means that the acquisition of non-mandated material is constrained by the goodwill of academics, and the staff time available for the ingestion of specific parts of the institution's collection, such as pre-mandated theses, or an individual researcher's profile of published items. The statistics each repository's website provides, allow depositors to see how many page views and downloads their items receive from all over the world. Usage statistics are currently the best tool coordinators have for promoting their repository. They allow depositors to find their work through keyword searching in databases and search engines. Statistics also help coordinators highlight the size of repository collections by faculty and identify popularly downloaded subject terms.

At this stage in OA history, there is an increasing readiness to do so across disciplines, once the researcher is aware of the repository's existence, but there are many factors that can inhibit this. In New Zealand, several of these factors stem from the comparatively recent advent of digital repositories reaching critical mass, around 2009-2010. Currently almost every university operates its

research department in parallel with its discovery group, and integrative software tools such as Symplectic Elements have been in use for less than a year.

Literature Review

The literature on IR indexing and visibility addresses two reasons for the reluctance to deposit in IRs: coverage, and copyright. Although coverage is limited by the resources available to LTS and metadata staff at each institution, the digital aggregation tools provided by DNZ: MET, MAT and WCT aim to bring the comparative accessibility of research in Google Scholar and DNZ member institutions closer together.

Depositor attitudes to Repositories

One of the longest running perspectives in IR publishing is users' views on how resources should be deposited, described, and displayed via an IRs API. A large part of this perspective has been explored in IR literature over the past four years, surveying depositors, subject librarians, and metadata staff. Some studies have utilised visual statistics to assess the international growth by subject in IR holdings since 2006³. Others have performed purposive interviews,⁴ and Likert surveys of depositors ranked by the stages of the academic lifecycle.⁵

The general trend discussed in such studies is the slow growth of subject holdings outside of mathematics and the sciences, where potential depositors are either oblivious to the existence of their repository, or reluctant to deposit again after the first three years of their career. The primary objective of IR coordinators in encouraging deposits, is building a collection that is visible on public search engines. This allows contributors to see their work being accessed by the global community, and see the repository as a free publishing system for their research profiles.

Scholarly submissions to repositories are described at the time of deposit by keywords and subject codes, as well as title, author, date and the other Dublin Core elements. The most important

contribution a depositor can make to enhance the document's visibility is to provide an adequate set of keywords that describe their work. In studies of self-archiving perspectives, such as those by Dickinson⁶, and Revell⁷, repository liaisons find that while depositors appreciate controlled vocabulary lists when describing their work, there is a strong preference for natural language.

Harvesting and Discovery of Repository Documents

There are three main reasons for this. Firstly, depositors use natural keywords when there are gaps or outdated terms in the controlled lists. Secondly, liaisons add their own natural keywords to submissions so that they associate better with related documents in the repository collection.

Thirdly, only academic databases use controlled vocabulary, whereas public search engines aggregate documents by natural language. This emphasises the need for international descriptive standards such as Dublin Core (DC) that are used by larger groups such as ANDS and WorldCat. By using DC as the primary list, coordinators can provide documents with handles that are more stable and accurate as the pool of OA material on the web expands.

This trend is highlighted in MAG 2012, placing DC descriptors above Marsden, ANZSRC, OECD, and PBRF subject codes in the contribution guidelines. This reflects two key issues for IR coordinators.

Firstly, although subject codes are useful at the national level, New Zealand lacks the technical oversight of an ANDS equivalent. This means that while PBRF codes are useful to research departments, Marsden, ANZSRC, and OECD description takes second place to DC, which is automatically visible to OAI-PMH harvesting. In smaller IRs especially,⁸ with less than one FTE staff member, there is simply no time to check and amend non DC codes added by depositors.

The second issue is the ongoing struggle of encouraging OA deposits outside of mathematics and the sciences. From a technical services perspective, the lower the quantity of OA full text items a given discipline has available on the web, the greater the amount of browsing which users must perform to find relevant material. To a depositor in the humanities, this may seem like a good reason to solely use prestige based publishers, but to an IR coordinator there are many other reasons for using

OA publishing as a second option. Perhaps the most cogent argument for choosing the Gold OA option is the increasing dominance of public indexing power over that of academic databases such as Scopus and Web of Science.

Since 2006, researchers have noted rapid developments in the literature on public indexing services. Although developments are arguably slow, given the quality of research resources made publicly available through Open Access agreements, it must be said that the contributions from academic consortia vary widely in quantity and accessibility. In Payne and Thelwell's (2007) review of 10 years of link changes, it is made plain that the largest hurdle to Google's aggregation of academic material has been the slow development of Open Access. This is due both to poor coverage description and to the Serials crisis of 2007-2009.

International coverage has improved dramatically since the crisis, as the number of DOAR (directory of open access repositories) affiliated groups continues to grow. This has also had the effect of cementing open source crosswalked platforms for electronic publishing as the standard partner in semantic RDF based storage. Cheap standardised platforms empower small institutions in relatively undeveloped countries such as New Zealand to grow beyond the pilot/implementation stage much more quickly. More importantly, widespread use by IRs of RDF E-print storehouses in semantic stacks such as Oracle, have drawn the federated ontology expressed in OAI-PMH DCMI standards far closer to the metasearch batch reformulations used in public indexing systems.⁹

Linked Data

The effects of this convergence have been noticed most strongly in link analysis studies. Most take a three faceted approach, analysing link source, target categorisation, and query reformulation.

Payne's (2007) review is part of an emerging body of proof that public searches for academic resources have routinely failed to navigate pre RDA ontology for years. A common conclusion of such studies is to call for a repackaging of encoding standards to provide searchers with more efficient reformulation options expressed through filters, suggestions and links.

“The most detailed academic interlinking motivation studies so far involve source and target page and link classification exercises using different categories for link creation motivation. Wilkinson et al. (2003) took a random collection of 414 links between UK academic institutions, downloaded both the source and the target pages and classified them according to the apparent motivation for their creation. They found that, by combining similar categories, more reliable ones were formed and, although less than 1 per cent of hyperlinks targeted formal scholarly publications such as journal articles or conference papers, over 90 per cent of targeted material was in some way related to research or other scholarly activity”.¹⁰

As the proportion of academic documents discoverable through public indexers grows, depositors and coordinators have an ever stronger incentive to use repositories as a direct-to-web publishing tool for institutional research.

Currently there are two main impediments to discovery of New Zealand’s academic information resources. The first stems from limited staffing in both the liaison and technical areas, and the resources LTS staff have available to put into the API and databases behind their IRs platform. In the literature, there are three main issues concerning information platforms and user interfaces in IRs, namely size, flexibility, and scope. IR development in New Zealand is primarily concerned with size. The subject indexing framework behind repository software such as DSpace and Greenstone used by NZ IRs relies on a certain range of user interface technologies to present content.

The guidelines provided in NRDS 1.0 and MAG2012 regulates the basic descriptive standards that allow users to retrieve resources by subject, author, title and date. While the simple nature of qualifying fields such as Marsden/ANZSRC or E-print abstract templates provide a good layer of accessibility to some types of content in New Zealand’s IRs, other types of information resources, such as conference materials embargoed post-prints and datasets, lack sufficient regulation in terms of indexing architecture. This results in the dichotomy between the results display indexed by public

engines such as Google Scholar or Baidu, and the repository interfaces of *research.digital.nz/DNZ* and its contributors.

Providing metadata for such items as datasets and conference papers is a major current issue in developing indexing architecture for repositories. Improving DNZ's and NZ's IRs ability to display such resources requires significant changes in both copyright and the ability of metadata harvesters to recognise a greater range of description of both subject and type.

Copyright Issues

The second impediment stems from copyright. Because Open Access and Institutional Repositories have been implemented in New Zealand only from 2006, there is a considerable fraction of material that has missed the initial digitisation for a number of reasons, most frequently because it is abstract only. For more recent material, even embargoed or pre-print items must have the author's permission to be stored in a restricted archive. By the time the resource is available IR staff must often use SHERPA RoMEO or OAKList to obtain final permission.

"A litany of problems plague current rights management processes—publishers' slow response time to author rights questions, overly aggressive licensing terms, unclear terms of licensing, and poor rights record-keeping."¹¹

Coordinators at New Zealand repositories react to these problems in different ways. One longstanding concern has been the unresponsiveness of SHERPA RoMEO. Combined with a lack of information on New Zealand publications, coordinators must either use Australia's OAKList as a close substitute, or compile their own lists for the eventual construction of a national rights database. There are many other issues as depicted in this graph from Hanlon's "Asking for Permission:"

Top Copyright Challenges (Survey Questions: “What are the top copyright challenges faced by your IR?”) ¹²

The second most common copyright issue IR coordinators face is providing accessible full text material via the licence that depositors sign before the research department passes a copy to the repository. This issue involves promoting the benefits of OA licencing in an environment where some depositors fear their research will either be locked, stolen, or abandoned in the repository. When collecting non-mandated material, IR coordinators must provide depositors with enough licencing options and evidence of the benefits of repository publishing through usage statistics.

Purpose of Study

This study adds NZ IR coordinators’ opinions on current developments in LTS coordination, and collection of non-mandated materials to the literature. Many empirical studies survey the collection policy of larger institutions with dedicated Technical Services and document management teams. Aside from Auckland University, most New Zealand institutions share technical support through a consortium, and operate completely separate research departments and library discovery teams.

Coordinators have no national data service such as ANDS, or copyright clearance directory such as OAKList supporting resource acquisition.

The extent of national communication lies in semi-annual IR Coordinator's meetings and mailing lists. By aggregating the opinions on how well the system currently works, and how the coordinators would like it to work the future, the study aims to provide a chronological snapshot of the NZ IR communities' development five years after inception.

Theoretical Design

This study reviews the common barriers that New Zealand repository coordinators face in collecting and providing access to their institution's research material. While most universities have had mandatory deposit policies for theses in place since 2007, the collection of other resources such as journal articles and conference papers relies on cooperation from authors, the research department which relays material to the repository coordinator, and the policy of the institution itself.

Given the unique situations of each institution, this study seeks to elicit the opinions of IR coordinators through an open and pragmatic dialogue in the form of a semi-structured interview. The opinions will be collected in an aggregated and unattributed manner to provide direction in a narrative analysis, using an interpretive constructivist philosophy. This approach by Strauss (1990)¹³ uses purposive sampling and grounded theory to perform an inductive data analysis. The interpretation is formed by the material trustworthiness of each coordinator's opinion. The opinions collected are quoted only if they are considered to be credible, transferable, dependable, and confirmable.¹⁴ This narrative approach is described by Cresswell 1994,¹⁵ where each participant's viewpoint is presented as an independent voice. This is important to the viewpoint this study takes in analysing the disparate situation of each university.

There are several examples where pragmatic realism is necessary. While some of the sampled coordinators may use similar repository technology and approaches in their roles in liaising with

researchers, the pragmatic approaches they take in prioritising their workload can be quite different. This can be due, for instance, to the workload balance created by staffing requirements, or the backlog of material available for ingestion. The framework assumes that each candidate is equally pragmatic, and their interactions as a coordinator with researchers at their institution and with colleagues at other institutions and consortial groups are driven by the same factors that drive other coordinators, other things being equal.

By collecting IR coordinators opinions on recent in developments in Open Access, the framework is used to construct a narrative list of factors influencing the growth in IR holdings, and the priorities of the coordinators. These factors and priorities are elicited through storytelling as described by Mishler (1986),¹⁶ where coordinators personal reactions to the constraints of their environment are added to the stories of other coordinators. These stories can be used to form a national narrative that is supported by the repository case studies and objectives produced by the NZ IR coordinators community in 2009 and 2011.

Research Objectives

The contemporary issues in New Zealand IR development are discussed in two documents:

- [The EDUCAUSE/CAIRSS summary](#) of Australian and New Zealand Repositories in 2009
- The 2/5/11 NZ Repository Coordinators Seminar

These documents describe the current situation and goals of each repository by institution, and by the common activities they share such as Metadata groups, which decide on common policy guidelines for OAI-PMH, and front end service delivery from the National Library. The annual seminars also discuss international repository activity, in terms of integrative technologies such as Symplectic Elements and OpenURL functionality, increased responsiveness from SHERPA RoMEO, and CAIRSS.

The main objective of the narrative analysis is to collect a cross section of the regular internal and external activity that forms NZ IR coordinators day to day such as:

1. Encouraging de-duplication of records from collection as PBRF records to markup with Dublin Core.
2. Attitudes on the value of Marsden, ANZSRC, and OECD subject codes compared to natural language.
3. Collection of grey material such as conference slides, journal preprints, and dark material such abstract only items, embargoed journal post prints, and theses.
4. Promotion of Open Access collection policy through the use of Creative Commons Licences, retrospective digitisation of items older.

All of these areas are approached within two contexts:

1. The coordinator's approach to these issues within his or her own institution.
2. The Institution's collaboration with wider groups such as LCONZ, National Library, CAIRSS.

The information gathered from interviews will construct a narrative of coordinators opinions that can be used in the analysis to evaluate the improvements in research workflows since 2009, when the sampled IRs consider themselves firmly established.

Research Questions

For Interviews:

1. How do IR coordinators prioritise the collection and digital uploading of non-mandated material?
2. How can IR staff assist depositors in describing content so that it forms a cohesive body of knowledge with similar resources discoverable through OAI-PMH harvesting?
3. How well do the national metadata guidelines represent the aggregated content of New Zealand's IRs?
4. Is the move to integrate the Institutional Research Department and Repository collection process improving depositors attitudes to OA publishing?

For holdings data:

How does the holdings structure at each repository represent the institution's:

1. RMS collection process in terms of usage agreements.
2. Metadata collection.
3. Proportion of dark material and grey material.

Research Design

This study interviewed the e-resources coordinators, or repository coordinators, who oversee the collection and ingestion of digital resources into their repositories. The sample included LCONZ

partner universities Waikato, Otago, AUT, and Unitec, as well as Lincoln, UC, and VUW. AU is excluded from the sample, since the size of its collection and repository team places it above the discussions for developing repositories.

The interviews construct a narrative of the parallel developments in New Zealand's Institutional Repositories. Each institution is in a markedly different position to its peers in terms of size, staffing, and holdings by discipline and content type. It was necessary to conduct interviews in a semi structured manner that incorporated the unique perspective of each coordinator into a narrative which reflects the collaborative effort between universities to grow each individual repository.

Each structured question in the interview was generalised to encourage coordinators to elaborate on their institution's position in a manner that can be compared with the positions of other institutions.

The issues arising from the narrative were contextualised by holdings data and metadata quality data. The data was taken from NZResearch's MET which collects live harvest data from each institution. The holdings data provides a background to each coordinator's discussion of the challenges in acquiring new and old material. This data is supplemented by each institution's metadata quality tables extracted by DNZ's MAT, and used to highlight gaps in the DCMI records for the institution's holdings data.

Methodology

Collection of interview data.

The interviews took place over a period of two weeks with the coordinators of eight institutional repositories: Unitec, Lincoln, Massey, University of Canterbury, Waikato, Otago, and AUT. Interviews with Victoria University of Wellington and DNZ were held earlier, in person. Research Space Auckland declined an interview on the grounds of being too established to answer the structured

questions. Unitec's Research Bank is the sole representative polytechnic repository since no response was received from CPIT, Wintec or the Open Polytechnic.

Using a semi-structured style based on the research questions, the participating coordinators' opinions were gathered on the main challenges in their role managing the institutional repository. The aggregated unattributed opinions were collected from each participant over the course of a one-off, half hour interview by telephone, recorded onto a computer. The structured questions were designed to encourage each participant to answer the research questions in the context of their own experiences. These experiences are presented in four sections to answer the four research objectives in terms of the individual and collective experiences of repository coordinators.

These experiences were then aggregated to form a loose narrative that can be embellished by the EDUCAUSE reports and 2011 NZ IR Coordinators meeting to construct a historical account of NZ IR development from 2009. The conclusion evaluates the NZ IR community's satisfaction with progress so far, and its hopes for the future.

The narrative was subjected to comparative analysis so that responses are subjectively weighted using axial coding. The main limitation of this approach is that each coordinator's experiences are based on the unique situation of their institution. The conclusions extracted from the transcript summary in the analysis can be compared only in terms of the repository's relative size and the nature of its relationships with wider groups. The coding system was used to weight responses to each scheduled question according to the narrative provided in the EDUCAUSE 2009 reports and 2011 meeting notes.

The transcripts are then quoted in the analysis to answer the research questions using the following open/axial coding structure from Strauss and Corbin (1998) ¹⁷ described by:

Base categories: IR coordinators opinions on:

- Liaising directly with depositors, and indirectly via metadata staff and research department staff.

- Maintaining the repository collection's metadata for subject mapping and discovery purposes.
- Maintaining an efficient repository interface with DSpace for searchers and depositors.
- IR Community support through LCONZ and the National Library's DNZ group.

Sub categories:

- **Why:** What options the coordinator has in this category area that forms their opinion. Was the decision made as part of a group, was it an experiment, or was there no other choice?
- **Who:** Which aspects of the repository's relationship with the research department or other institutions made the coordinator choose to act in a particular way. Is the coordinator's aim in this area to improve a part of the relationship in question, or to end it?
- **When/What/How:** How long has the situation being discussed being going for? How has it impacted on the IR staff's workload? How is the decision expected to improve the repository's service delivery or functionality.

Collection of Holdings Data

Holdings Data was collected from NZResearch, aggregating the filters into the following groups.

- Thesis.
- Conference Item / Conference Paper / Conference Poster / Conference Proceedings/ Working Or Discussion Paper.
- Journal Item / Journal Article / Submitted Journal Article.
- Book / Book Item / Book Review / News Item / Patent Report.

The data is presented in a table by institution for the following date ranges in which the items of each group were published:

1900 - 1999
2000 - 2009
2010 - 2012

Although each institution provides holdings data publicly, it is not presented consistently across the XMLUI behind interfaces. This is because of the variety of DSpace statistics add-ons and configurations in use. There is a mix of information available from NZ IRs including Google Analytics retrieval data by browser history and country. There is also retrieval data of popular subject and keyword searches.

The data is presented in bar charts to illustrate the coverage of each repository's collection by date and type. An important limitation of this data is that some institutions lack complete DC records in

many parts of their collections. There are many reasons for this, most commonly because the records were never given that information when they were added by the research department. When such batches of records were later ingested by the IR, metadata staff lacked either the time or the information necessary to complete the records before the bulk upload.

The NZResearch holdings data was checked against the institution's copy. In cases where there is a serious flaw in the DC records, the institution's copy was then used in place of NZResearch data. The challenges in the data collection are:

No date identifier: In this situation the holdings data by type is added to the 2000-2009 group since the initial uploading period ran from 2006-2009.

No format identifier: In this situation only identified records are included in the type lists, unidentified records fall into the 'other' category that is represented only in the institution's grand total count. In some institutions this includes the contents of the restricted archive such as abstract only items.

Broken NZResearch filter: An outstanding NZResearch bug of 2012 is the inability during a single search to switch filters between institutions, type, or date range. The data collection was repeated three times during the research period to eliminate any error.

Limitations

Collection of holdings data

The study aims to show holdings by type represented in three periods of publication, to provide the following information:

1900-1999: the issues involved in collecting back files.

2000-2009: The proportion of holdings by type in repositories uploaded in the first implementation in 2006-2009.

2010-2012: The coordinators success in collecting non-mandated material.

As described in the methodology, the data was collected from NZResearch. The process was triple checked to prevent simple filter errors, but the main limitation in this process is the proportion of unknown records. It is impossible to tell from NZResearch whether the documents are in a restricted archive or a dark archive. It is also difficult to be certain that the total counts are accurate. DNZ has

limited development time, and coordinators have encountered significant harvesting errors as more IRs were added. The data is cross checked against the public statistics of each institution, but it is necessary to limit the findings of this area of research to the generalisations made by coordinators about their collections.

Collection of Interview Data

IR coordinators come from different backgrounds in terms of their previous experience and current responsibilities. Institution repositories have different collections in terms of discipline structure, the time since they were established, and their relationship with the R.D and depositors. The interviews attempted to combine these threads to accurately portray each coordinator's interpretation of the challenges to repositories at the institutional and national level. The answers given to the structured questions were used to ask follow up questions using the subcategories in the axial coding design.

Aggregating responses from so many different perspectives has made it necessary to keep the narrative built by quotes in the analysis extremely general. This process assumes that each coordinator is equally rational and would give the same responses at any hypothetical IR of comparable size and setup.

These responses are used to form a historical narrative of the last five years of IR development. The supporting context of the proportional holdings data at each institution is intended only to show the diversity of publishing dates and formats across IRs.

The final limitation is that only Unitec is sampled among polytechnic repositories. The main reason for this is that no response was gained from Wintec, Open Polytechnic, or CPIT, but it does allow the study to focus on DSpace repositories.

Analysis

This section discusses the key issues raised in the interviews in terms of the research questions. Excerpts from the interviews are used to illustrate the issues central to current IR developments. Each excerpt is a contribution to the narrative discussing the common barriers and improvements in collecting and publishing resources in IRs. Bar graphs displaying NZ IR holdings counts are used to show the proportion and recentness of each type of information resource in NZ IR holdings. Each repository coordinator expressed varying levels of satisfaction at the progress of their institution. The analysis of the answers given to the research questions builds a narrative which focuses on the common issues in an aggregated form. This approach assumes that in terms of national policy methods of collecting and publishing institutional research in repositories, each respondent is equally rational.

The aggregation does not extend to policy at the level of individual institutions. The perspectives collected on issues such as full OA collection, and relationships between repositories and research departments are presented as separate fragments. These unique perspectives are added to the narrative defined in the research objectives through the axial coding system of categorisation. Each response is given a qualitative worth in the analysis by relating it to the national trends. For example, in Part 4, the issues raised by coordinators of institutions where integration has not yet begun is compared with the experiences of coordinators where integration is complete. The narrative analysis process is entirely subjective - because each institution is independent, each answer is a unique reflection on the contemporary problems and barriers in building an IR's collection.

In the conclusion, these reflections are drawn together in the context of the common goals outlined in the EDUCAUSE/CAIRSS 2009 repository case studies. The 2/5/11 IR Coordinators meeting is used as a qualifying framework for the reflections generated by the structured interview questions. The perspectives given by coordinators in reaction to their IR's situation is related to the group discussion in the 2/5/11 meeting. The conclusion discusses how well the relationship between

coordinators and their repository is supported by the New Zealand and Australian IR communities.

This discussion covers three areas outlined in the research objectives, and structured using the research questions:

1. How do IR coordinators prioritise the collection and digital uploading of non-mandated material?
2. How can IR staff assist depositors by describing content so that it forms a cohesive body of knowledge with similar resources discoverable through OAI-PMH harvesting?
3. How well do the national metadata guidelines represent the aggregated content of New Zealand's IRs?
4. Is the move to integrate the Institutional Research Department and Repository collection process improving depositors' attitudes to OA publishing?

In this section, the analysis forms a narrative pattern from coordinators responses to the research questions. The narrative description in each area is supported by quotes describing the most common issues coordinators face, and is extended by specific problems that are unique to individual institutions.

1. How do IR coordinators prioritise the collection and digital uploading of non-mandated material?

During the initial rounds of digitisation between 2006 and 2009, NZ repositories uploaded the bulk of theses collected before and after the mandate. These were mainly published in the 2000 - 2009 era, as can be seen in Chart 1. Most institutions digitised theses in bulk, adding between a third and a half to their restricted archive. The official number of abstract only theses kept in dark archives is less than one hundred at most institutions, but because of the separation between the research department and the repository, the real number of these theses may be a matter of several hundred.

Chart 1

As Chart 2 shows, the University of Auckland repository holdings dwarf those of the other New Zealand Institutions. Research Space at Auckland hosts a comprehensive array across the four categories of digital resources with and without full text. Other IRs have varying arrays available according to the time constraints of staff, and their relations with depositors and research coordinators. They also use different methods of presenting the information to browsers depending on the IR's Manakin interface configuration and statistics package.

Each institution's holdings represent the collection pattern of their coordinator's activities. These patterns are reflected in the MAT statistics collected by DNZ. For example, Victoria University of Wellington's collection includes thesis material from its open and restricted archive, as well as journal articles, technical, working and conference papers, but they are not identified in Metadata Extraction Tool (MET) by type. Chart 2 reflects only the raw number of the records which MET

counts as visible, complete metadata records. Consequently, in Chart 2, all of VUW’s holdings have been placed in the theses category, as that is what the majority are likely to be.

Chart 2

The other major problem with the data is reflected in the 60% average quality most institutions score in the Table in Appendix F. As well not harvesting records with individual fields, DNZ misses material in dark archives such as abstract only records. There is no fixed pattern for keeping track of all the material even in the unrestricted archives. Research Space at Auckland for example has only 1,000 out of 14,000 records visible at the time of writing this report.

Chart 3

Since the critical mass stage was reached in 2009, when IRs formed a stable relationship with DSpace and their research department, deposits of non-mandated material published since 2000 have grown to represent between one fifth and one quarter of NZ IR holdings at 5 out of 8 IRs. Chart 3 demonstrates the increasing proportion of post 2009 Journals, Conference items and Book deposits in IR holdings. This is a reflection of the improvements in coordination between the repository and research department, allowing IR coordinators to operate on a 'steady as she goes' basis, focusing their attention on collecting material missed in the initial digitisation rounds of 2006 - 2009.

For material published before IR coordination work began, the collection process is limited by the extent of full text availability. A common problem is that hundreds of records are abstract only. Such records often also lack description beyond PBRF identifiers. This means coordinators are doubly thwarted and must rely on contacting the authors to obtain the full text. This can be seen in Chart 3, in the low proportion of non-mandated 1900-1999 material holdings.

Collecting this group of material involves two main challenges to coordinators. The first is obtaining permission to digitise. This may simply be a case of contacting the author(s) and saying 'hey it's not on, do you want it on?' as is often the situation with conference material. With books, journals, and theses published before 2000, many have had copyright strings attached at the time of deposit to the R.D resulting in a cache of abstract only items. Coordinators had differing views on such caches:

- "We have a very small number of embargoed non OA material, which drives me batty - I can't see the point of having restricted stuff you can't access."
- "I'd quite like to batch upload older theses, but we haven't had the funding, but we haven't been going that long so the majority are online anyway."

That's the theory of Open Access, that just putting something online can only increase its chances of being read?

"Exactly, with retrospective digitisation of theses older than five or six years we thought it's (copyright) probably not going to be much of an issue."

These views demonstrate frustration with the difficulty of obtaining back files, and stressed the importance of having an effective relationship with the R.D. Since NZ IRs have been operating for six years now, many of the problems in the collection policy of R.Ds have been solved. There is a range of issues that will be further discussed in Section 4 of the analysis. In terms of prioritising collections, the main theme given in the responses was that collecting material from the R.D is most efficient when each stage is automated, from the collection of OA rights clearance to acquiring the full text record with all of its PBRF and DC metadata intact.

2. How can IR staff assist depositors by describing content so that it forms a cohesive body of knowledge with similar resources discoverable through OAI-PMH harvesting?

When a document is passed from the research department to a repository, metadata staff check that it is supplied with adequate descriptive information to enable it to become a discoverable digital object. The most basic set of information is the PBRF and DC data, describing the author(s),

title, and date, as well as the publisher, format, and licence information. Uploading this data is a largely automated process of copy cataloguing for metadata staff and is often done as a bulk ingest through the DSpace software. This includes adding the handle that is used to partner the DOI in creating a stable internet address for the document, which can be embedded in the depositor's profile page, and the repository's index.

The chief manual task of IR metadata staff uploading new submissions is checking the subject codes and keywords with which the depositor and research department have endowed the document. The purpose of this task is to maintain a structured body of research that can be indexed and harvested through the simple crawl specifications of public engines such as Google, Baidu, and Bing .

Depositors are encouraged to apply subject codes and natural language keywords from encyclopaedias provided by the research department and repository.

The area of interest in this research question is what value IR coordinators currently place on these codes and keywords. IR coordinators have two responsibilities in this area. Firstly, they must work with LTS staff to maintain the metadata quality of their collection for harvesting. Secondly, they must ensure that documents are described so that they join related documents in a filtered search of the Institution's database, and other academic databases. The main issue this question explores is if the growing dominance of public search engines over databases is encouraging the abandonment of official coding methods such as OECD, PBRF, Marsden, and ANZSRC, in favour of natural language keywords. Both keyword and subject codes are equally discoverable by an OAI-PMH harvester, but not every repository provides documents with controlled vocabulary such as subject codes. The most common reason for this is that there is no staff time available to mark up documents with controlled terms, especially in Green OA repositories such as AUT which supply only OA full text material to the web.

The responses quoted in this section are typically validated by the coordinator's knowledge of the institution's Google Analytics data and DSpace statistics. This data displays:

1. The proportion of new versus returning visitors to the repository.
2. How many documents they viewed.
3. Which country they are from.
4. Which site accessed the document handle e.g. Google Scholar, or a helicon address from a library database.

A typical percentage breakdown of this data shows that between 60% and 80% of visitors are new, and view only one or two records. Around 30% of visitors view three to six, and 15% view six or more before leaving the repository website. In each repository 60 - 65% of access requests come from Google or another public search engine, and a further 10 - 15% come from follow up requests, i.e. the user browsing a related record using the repository interface, having discovered it through Google. Finally, approximately half of the hits a resource receives come from a New Zealand city, and of that usually around 80% from the institution's home city. A quarter comes from the U.S.A. and Europe, and the remainder from Asia. These numbers exclude robot harvesters which account for 30% of hits.

After discussing usage and retrieval statistics, the interview candidate was asked about their perception of the benefit of controlled vocabulary. This includes the Marsden research fields, OECD and PBRF codes, and the ANZSRC system which has recently been implemented at some institutions to supplement and supersede the Marsden codes. All of these codes have decreased in importance in the MAG as a result of discussions from the 2/5/11 IR Coordinator's meeting. At the same time the importance of completeness in DCMI records has increased.

Sample excerpts:

So how much effort goes into making sure records have accurate metadata, for example Marsden codes and ANZSRC?

"We don't use ANZSRC at this time. Everything that's deposited gets looked at by our department, information resources, which looks at the data input, purchasing material, and the staff in there actually set the process up, check the records, check the file is not corrupted, and then push it through, mapping it into the department collection. The checking is important when theses come

through from the department for graduation, they also check for doctoral theses, that the supervisor and the title matches.”

Is there any benefit to these codes do you think?

“We had the Marsden codes in a previous version of DSpace, and when we upgraded we lost them, and we’ve never implemented them again, and we’ve not gone to ANZSRC either, because it is basically not a priority for us.”

“We do try to ensure that each item that goes in has at least one Marsden or ANZSRC code, the later ones have two, sometimes three depending on the subject area. The more specialist things are we tend to offset the lack of a Marsden or ANZSRC code (from the depositor) with a couple of extra keywords from the main keywords field so we’re covered from all directions possible.”

Why did you choose ANZSRC?

“Well we were using Marsden with BePress when we were part of CODA with the other polytechnics, and when the funding ran out, we moved to DSpace and ANZSRC seemed like the up-to-date thing to do, I didn’t really pay much attention to what the others were doing.”

What’s your position on controlled vocabulary?

- “There was discussion on ANZSRC when AU applied it, but they also use DC and we’re quite happy with that, it’s still being updated”
- “When people deposit things in our database we do provide those codes in a thesaurus for them to use, but it’s optional really, generally they just use the natural language keywords.”
- “We don’t apply those codes anymore, the retrieval comes from Google keyword searches rather than database searches”

Opinions on the value of subject codes vary chiefly by the amount of staff time available to apply additional subject codes. This factor was often qualified by the coordinator’s personal opinions on the value of subject mapping in building an index for the repository, where faculty research is represented by cohesive groups of linked items. Another problem suggested was that subject codes are no better than keywords: if there are too few, an item can fall into a cul-de-sac. Dickinson finds

depositors frequently select keywords from the title or abstract of a document, and find using thesauri too difficult.¹⁸ The coordinator's opinions on this typically fell into two categories.

Coordinators of OA only repositories favoured purely natural language searching through Google. These repositories are characterised by the recency of their collections and the institution's emphasis on technological disciplines in its curriculum. Depositors in Mathematics, the Sciences, and Technological disciplines were frequently cited as being the greatest supporters of OA publishing. This position was linked to an increased affinity gained from using OA resources provided by subject repositories serving these disciplines. Because OA repositories provide resource description through standard DCMI OAI-PMH, searchers and depositors were implied to be greater consumers of Google as a first search option, as opposed to databases using controlled vocabulary.

3. How well do the national metadata guidelines represent the aggregated content of New Zealand's IRs?

Despite its limited capacity, DNZ provides a valuable metadata service. The metadata dashboard highlights which areas of each institution's holdings lack full DC coverage in areas such as date, format, and rights. This information can be used to illustrate the problems IR coordinators face in collecting incomplete records from the R.D during the ingestion of back files or previously embargoed material. In general, the Average Metadata Quality Tables in Appendix F indicates how far each institution is from providing perfect metadata records, depending on the proportion of incomplete records in its holdings. The Tables in Appendix F demonstrate the variety of challenges DNZ faces in building a national index. Some institutions lack 50%-100% of important technical DC fields such as e-prints, handle, and contributor.

At the 2/5/11 IR meeting, coordinators formed four groups to discuss how KRIS/DNZ could help improve its metadata harvesting and aggregation process. The Shared Search Infrastructure through which DNZ collects metadata from NZ IRs can be browsed by type, date, and author. Since the

updated MAG was released in March 2012, DNZ has added multiple author and rights filters to its portal. As the Appendix F Tables indicate, about half of non-mandated items lack searchable date or type fields. This is due to the problems discussed earlier caused by collecting incomplete records from R.D. The consensus at the meeting was that the barrier should be intentionally lowered so that incomplete records could still add to the national harvesting totals, improving the visibility of NZ's research communities.

- **Do you feel the guidelines that came out this March are the best thing that could have come out of those (2/5/11 NZ IR Coordinators) meetings?**

"I contributed to those guidelines. They were the best we could achieve at that time, and they needed to be updated. They're not perfect, I think the biggest outcome is moving KRIS from the original system to the DNZ infrastructure, not saying it's perfect now, but it means you've got a large pool of technical specialists that still exist in DIA, now they've moved it to a system where they have the experts that DIA is willing to keep investing in, as a front end to everything, and I think, although it's not perfect, I'm actually quite happy with what they developed in the interface, and they have been responsive on some issues. Things like being able to filter by Creative Commons now, so we've now figured out how to do it, it means some of our stuff comes up as being used commercially, so it's a way to promote these researchers who want to share these outputs openly. It can be quite powerful, because we're not sure how successful Google is with their CC licenses, so it's good that within NZ we can find and focus on that type of material."

Indexing CC fields consistently across NZ IR's remains a major obstacle. At this stage, DNZ's priority is adding an OA full text field, as a placeholder for a consistent rights field. Another major issue the MAG attempt to address is the disparity in subject and keyword indexing across IRs. Lowering the barrier has allowed smaller IRs to contribute purely DC based Green OA records to NZResearch OAI-PMH harvests, but until it gains more development time DNZ is unable to provide subject filtering.

There are plans to launch a prototype version of SKOS, that the ANDS is currently preparing a Q4 release for CAIRSS members, but there will be a gap of about two years before DNZ can add a subject mapping framework that effectively services all NZ IRs.

Would you like a national data repository?

“I think it’s not on a lot of people’s radars yet.”

➤ **Don’t they know about the idea?**

“Well, where is it going to start from? It’s in the department, there’s meetings going on and we know it’s a world-wide issue, but it’s about scalability, the library is somewhat involved. We don’t have something like ANDS, big Australian coordination, everyone’s doing their own thing, It’s a secret, like everyone is trying to be first. If we had that centralised help like Australia has with ANDS, you can see them getting ahead in leaps and bounds in so many areas.”

- “The beauty of an OA repository that is OAI-PMH compliant, is people can key in what they want without having to think about where in the world the information is coming from, so personally I don’t see what a national data repository could be used for. Oh you mean making it cloud storage, who’s going to be the front, the National Library? The storage part would be good, but the National Library interface would be redundant because, you already have the repository interface.”

When asked about their position on subject mapping, coordinators at larger institutions with wider discipline ranges in their collection tended to prefer applying controlled vocabulary when necessary. The main rationale for this was that existing material in the collections use the codes, and there is still the existing relationship between Marsden/ANZSRC and PBRF coding at the point of deposit.

Two candidates mentioned the importance of building a national subject index such as the ANBD, which represents New Zealand’s unique research. In the context of national collaboration at NZ IR

meetings, the candidates stated such an index is currently the sole work of cataloguing experts at individual institutions, and is frequently dropped as the experts leave. Several candidates confirmed a desire by the IR community to have a national data service provide a national index, such as that provided by the ANDS. It was also acknowledged that DNZ at the National Library are currently limited to harvesting IR material 'as is', having lost Oracle access in 2011.

4. Is the move to integrate the Institutional Research Department and Repository collection process improving depositors' attitudes to OA publishing?

The main point made by each candidate in regard to working with the R.D is that the repository's role is as a storage space. The coordinators promote CC licences to R.D staff assisting depositors acquire permission for publishing older material, or grey material. These activities take time, and rely on the cooperation of authors, and the quality of material held by the R.D. These holdings can be transitory. The process of building a non-mandated collection of refereed and licenced material with full text attached requires coordinators to maintain an efficient relationship with R.D clients. The IR coordinators purpose is to encourage repeat deposits by touting the IR to authors both as a publishing and a preservation tool that does not encroach on their rights.

In obtaining OA rights clearance, coordinators use the international and Australian OA licensing databases SHERPA RoMEO and OAKList for the majority of their work. The issue in this area is that New Zealand lacks its own national database, meaning that coordinators must continually contact Sherpa, and also keep their own lists when necessary. At the 2/5/11 meeting, a national list was discussed but not implemented. It was felt that since 87% of NZ IRs use OAKList, the coordinators were better simply managing their own lists as necessary.

Do you think Sherpa's comprehensive enough, do you have a list of your own?

"We don't have time to make our own list, But Sherpa's like DC, they've pulled themselves up, because two years ago when you emailed them saying they didn't have this journal on their database they'd just ignore you. But now they're very reactive to the users."

Is there any collection policy for encouraging non mandatory deposit?

“I guess we’re a bit behind the eight ball here because we’re just getting our project to integrate the RM tool Symplectic with DSpace, and with PBRF, over the last 18 months because it’s all been about PBRF, it’s been about getting the actual records rather than the full text. Now that we’ve got through that stage, our next mission is to get not only those records into the repository, but also the full text.”

➤ How recent are these records?

“Certainly the last six years. Once that project is up and running we’ll be doing a lot more publicity to get the full text contributed to the repository. Whether that becomes mandated will be up to RM services, we don’t have that authority, we’re just a storage space, a discovery service.”

The interviews revealed a certain level of stoicism mixed with caution regarding the fate of these records. The most compelling reason one coordinator provided for chasing older non-mandated material was for expanding currently topical areas. It was implied this would promote retrieval of related information, and improve the goodwill for deposit from that research community.

- **How long ago were these restricted items published? Do they have an embargo period? Of a year for example?**

“Well we mandated these from January 2008, we don’t mandate dissertations because they are generally deposited by supervisors and students. The material that we put in retrospectively dates from 1928 at the earliest, and some of the highest used ones are from the late 70’s, early 80’s, where it’s probably resource management issues that are coming up, there’s Treaty of Waitangi issues, all those sort of things that are coming round on a cycle again today, that had been dropping off through the 90’s. But now overseas people are picking up on some of those issues.”

Another common attitude was that goodwill can be acquired retrospectively by publishing whatever material was available, including preprints and grey material. This preservation philosophy hinges on

the repository simply acting as a publishing tool for researchers, relying on the research department for quality control.

Have there been efforts to get back files if the author offers them?

“The focus has been more about awareness of Open Access rather than pushing the deposit of their research into the archive. So things like discussion papers, the grey literature, and when the department no longer have an electronic copy, the research archive will have one for them - - - so we have it in our archive just to try and increase the impact of the published version. So the repository’s a good way to push up results, promote articles that are not in high impact journals. There’s not a concentrated effort but there’s certainly things we have like the scholarly communications guide which is about more than us (the repository), it’s about publishing and access, and data management.”

A related theme that came up in this area was that it is the repository’s role to engage with willing contributors to publish all the material they wished beyond the initial deposit. Several opinions supported the viewpoint that publishing everything available to the web through repositories was a purely positive way for institutions to support their research community. This viewpoint was qualified by the need for preserving an item before it fades from the R.D’s notice, and also by the benefit of increasing the item’s presence in the international online community. The preservation argument for grey literature can be extended to pre-print journal articles, leading to an increase in post 2009 journal deposits.

Sample Excerpts

- “Anything in the RMS I regard as fair game, whether I can actually get access to the papers is another matter, conference papers are usually online so we download a copy before the link dies.”
- “there is the attitude that the final version is the quality version and ‘I don’t want people to see the preprint’. The culture is definitely different across disciplines: sciences, engineering,

computer science are open to post prints, but with humanities, there's a lot of other materials to publish, for example it's really hard to put a book online."

- "Usually they're cooperative, they just didn't realise we don't allow abstract only. It's only a small percentage that have a 'take or leave it' perspective. There's a refreshing trend towards post-prints. Last year there seems to be a considerable increase of published items from academic staff, and publishers from OA journals. "

- **How do you deal with reluctance to deposit?**

"With us it's not such a big issue, because my time is limited (as the sole 0.5 FTE) so I basically grab what I can get, and we've got enough staff members who are keen to keep me busy. So mandatory deposit is a bit of a non-issue. We're doing quite well, it would be nice to have more staff time to attack growing the collection, what we don't have is the automated deposit system."

The main response to this question across institutions stressed the importance of collecting material when it is first available in the RD. By promoting the benefits of repository publishing to depositors with usage statistics, and providing a free choice of depositing embargoed material in a restricted or dark archive, coordinators with sufficient staff are able to save time later on gathering permissions and full text. There was also a positive perception overall about the IR's influence in promoting OA publishing. Statements such as 'grab what I can get', 'steady as she goes' illustrate the limited time available for collecting back files from the research department. In 75% of the interviews, coordinators stressed the importance of an efficient relationship between the institution's repository and research department, and the need for an automated transfer of full text documents with complete metadata using a protocol such as HERDC¹⁹ or SWORD²⁰.

A fundamental problem in aggregating opinions on this issue is that the sample institutions' R.Ds have very different collection methods in terms of keeping an up-to-date PBRF database and consistent rights and descriptive information. It was unclear how many potentially items repositories fail to collect, but of the known unknowns that coordinators add to restricted or dark archives, the numbers given ranged from six to a hundred. At this stage only the University of Auckland and a couple of LCONZ members have implemented Symplectic for the integration of the repository and RMS. Integration provides clear improvements to self-deposit as can be seen in Auckland's collection, but cost is not the main problem.

- "I had thought at CAIRSS we were far behind, in 2011 they were only talking about transferring the RMS data from the dark archive, and now we've actually done it before them. Now researchers, once we've approved the manuscript, can publish straight to their profile, whichever version they want."

Do you work with others on DSpace?

"No, we've had to develop our expertise which has taken quite a long time, which is why we don't have as many customisations, we do have a training session in November. I don't think you need to be in a formal group to do that, other than that, we get requests for technical stuff, I try to help as much as I can."

LCONZ potentially has funding and technical expertise to implement Symplectic once it has been trialled, but the more important issue the interviews raised is the organisation of the R.D itself.

How well does your RMS encourage people to use shorter embargos and allow post prints?

"Its not very integrated at all, we're quite manual. Because TEC wants the final author's version, it actually has been more of a hindrance if anything. Theses are really easy because we've had a mandate in place for years that works really well."

Ah, that was my other question on RMS, at other institutions the research department gets the first copy for PBRF purposes, then sends it the repository which marks it up with some more metadata. Are you saying at your institution it's the other way round?

"Well in our case we don't yet have deposit for PBRF, though we certainly use the archive for PBRF, we encourage staff to provide a copy of what they're allowed to deposit. In some case we miss out, but where we were able to retain a copy and archive it, we did that, and then provide a handle or DOI back to the PBRF process, so we've probably done about 3000 links for them overall, and quite a lot of that we've been able to retain and archive, and we're working on a process for the future where we're first in line to receive the content."

So are there any problems with having a separate Research Department and Repository, for example duplicated or insufficient metadata?

"Not really, because we don't have anything concrete, and it's not just the research system, it's also the student system, the university is putting in a new student system, which will include the postgraduates, and the theses and the supervisors. So it's up to me, as someone coordinating developments, to be aware of these possibilities, but it's got to be a university drive to do it, and then for the library ready to support this."

The answers given to this research question show that successful integration depends more on the coordinator's ability to promote the repository across the university, rather than simply implementing a system such as SWORD or Symplectic when expertise and funds allow. The efficiencies created by streamlining self-deposit certainly encourage continued deposit by researchers, and lower the time costs to R.D and I.R staff in storing deposits²¹.

It is also the consensus that the difficulties of collecting full text permission for incomplete records are made considerably worse by the separation between the R.D and IR. The main pattern that emerged in response to this question was that the IR is limited by its role as an archive and discovery service for the R.D. Although coordinators can encourage OA publishing through the R.D, there is still

an ingrained resistance in some parts of the research community to depositing more than PBRF data to the university.

Previous studies of NZ IR depositor perspectives by Dickinson, Chawner, Reid, and Revell have all found that the variety of reasons discouraging deposit, result in the perception of IRs as a storage space mainly for theses. The overall coordinators response to this perception was that developing coordination between the IR, the R.D, and depositors, generally promotes the perception of repositories as discovery services, thus increasing goodwill. Only half a dozen instances were mentioned at each institution where documents were kept as restricted abstracts. By providing a full range of CC options, and using policies such as full text only, the repository can maximise the range of material it can obtain from the R.D in whatever form the depositors prefer.

Conclusion

Since the NZ IR community reached critical mass of 1000 records, coordinators have operated on a 'steady as she goes' basis. Beginning with seed funding from TEC, CODA, and LCONZ, most IRs have successfully used DSpace for 3-5 years, sharing technical advice, and collaborating on national metadata policy for harvesting. As described in the MAG, NZ IRs have the basic infrastructure to provide discoverable records with near perfect handle rates, and average DC scores of 60-70%.

Integration and OA

As the infrastructure develops, coordinators expect integration to increase deposit rates as researchers become empowered to publish items directly to their academic profiles via the repository software. Much of the workload faced by depositors, and IR/R.D staff in publishing IR material, depends on the relationship between the IR and the R.D, whether it is forwards, backwards, or single entry. There is a powerful incentive to choose OA IR publishing as a fast and effective way of contributing research to the online community of the depositor's choice. The holdings data collected in this study show a definite increase in deposits of non-mandated material since 2009

when IRs reached critical mass. At the very least, this demonstrates that IR's are successfully collecting recently published conference items and journal articles, respecting embargos on full text when necessary.

Because of the limited staffing at each IR, there was no expectation of any dramatic events, such as a global shift to Green OA, affecting NZ IRs. The longstanding resistance to repositories in some research disciplines, and the opposing affinity for them in other disciplines, was considered to be a non-issue. The 'take it or leave it' attitude remains a factor over which coordinators have no control. By providing publishing options from Green OA to Restricted Abstract, coordinators can at least build some awareness of the IRs role as an archive, if not as a publishing tool.

Prioritising Collection Building

The prospects for increasing digitisation of grey literature and back files are mixed. The main problem is acquiring full text, rather than permission to digitise. Again, staff time is a limitation. A common perspective was that only material that would provide an immediate benefit to the collection should be given priority in the bulk ingest queue. Potential benefits discussed include expanding the holdings on a currently topical issue, or the opening of a new sub-discipline representing a faculty or department's collection.

Phrases used in discussing grey literature - 'more the merrier', and 'grab what I can get', illustrate the perspective that there is a place for all academic material in repositories. Four coordinators stated that repositories were useful tools for promoting otherwise unpublished material, and building a public digital presence for research communities within the institution.

Perspectives on subject mapping and public indexing

The analysis of integration, and the barriers to collecting non mandated material have contributed to a fairly uniform perspective amongst IR coordinators. Half of the sample did not support the use of Marsden or ANZSRC in preference to DC. This was firstly because of depositors' general preference

for natural language, and secondly, because of the staff time needed to apply supplementary codes. The other half of the coordinators supported the use of subject codes as a form of controlled vocabulary representing Australasian research.

One perspective that was found across both groups was that public indexing through Google is perfectly adequate for discovering natural language DC records and that subject codes are not necessary. This view was expressed most strongly at the smaller institutions, providing nearly 100% full text collections from 2000 onwards. At larger institutions with older collections that have used Marsden classification throughout the 20th century, subject codes were perceived to be part of the indexing structure. Coordinators with this opinion pointed out that their institution's research outputs have been clustered using the codes for so long that they are an embedded part of each faculty's history.

National IR perspectives on OAI-PMH aggregation, and data service support

In the final section of the interviews, coordinators were asked how satisfied they were with the national level of IR coordination, and their expectations of improvements in the future. Since the general implementation of DSpace 1.6 to the current 1.7.2, the NZ IR community has enjoyed a period of steady liaison work with occasional projects such as backfile ingestion. Each IR Coordinators meeting leading to the drafting and publishing of the MAG, has stressed the importance of lowering the entry barrier for IRs to provide harvestable material to the web. At the 2/5/11 meeting, it was agreed that DNZ's role, given its limited resources, should be a light front end. Using MET, MAT, and WCT, DNZ acts as a moderator of NZ IR harvesting, monitoring metadata quality, and acting as a tide mark for the progress of each IR towards Green OA.

Currently NZResearch is better at showing which records are poorly visible in OAI-PMH aggregates. In the Table in Appendix E, it can be seen that many institutions have between several hundred and thousands of unknown records. The most critical fields absent in such records are format, access, and date. At some institutions the records are simply in a restricted archive, or abstract only. At

others, the field was never collected, or is unrecognisable. NZResearch currently sees only 1,000 out of Auckland University's 14,000 records, for example. Coordinators acknowledged during the interviews that DNZ has a big job and has done well given its limited project time.

The study found considerable cynicism with regard to the DNZ's ability to help, shared by IR coordinators and DNZ staff. Most repository coordinators conveyed a strong attitude of self-reliance, especially at non LCONZ institutions. When questioned about hopes for the future, the possibility of producing a national data index was the central issue. ANDS was mentioned repeatedly as an example of the type of national coordination IRs need to support technical projects such as adding DOIs, which could ultimately produce a linked data repository.

Some National Library work in RDA is making steps towards this, but the strongest narrative theme in this area, is that the best help in terms of advice is currently coming through CAIRSS. Several coordinators stressed that because New Zealand's IRs are so much smaller than Australia's, there are different priorities in their technical services requirements, and this allows a degree of experimentation with data management on a small scale. For example, the conversion of dark archive material directly into full text documents on a user's profile in the form they prefer, is a considerable step towards encouraging OA deposit. Successful experiments encourage partner IRs to develop similar systems that suit their specifications, and the further such technology spreads, the better repositories will be appreciated as publishing and discovery tools.

Suggestions for further research

This study set out to examine the age and diversity of NZ IR holdings and the challenges coordinators face in managing those holdings and promoting deposit. It has found that the barrier to collecting material varies according to the IR's relationship with the R.D's workflow. Coordinators primary obligation is to collect what they can, and the Holdings Charts show encouraging increases in non-mandated deposit.

A useful continuation of this research would be to examine the increase in non-mandated deposits by schools, if not faculties. If a consistent sample population could be found to analyse, reluctance to deposit may be re-evaluated beyond 2012.

Another area worth investigating is the benefits of CAIRSS membership for NZ IRs. Each repository uses its own configuration of data management policies and software. There is no unified setup for data management in New Zealand, and tools like SWORD and Symplectic can only do so much to bridge the gap between IR and RMS workflows. A narrative study into the history of CAIRSS's work with smaller IRs could provide insight into how coordinators can contribute to the eventual development of a national data service in New Zealand.

Glossary

Aggregation: Grouping harvested material by metadata type or encoding.

ANBD: Australian National Bibliographic Database.

ANDS: Australian National Data Service.

ANZSRC: Australia New Zealand Standard Research Classification.

API: Application Programming Interface.

Axial Coding: The grounded theory process of disaggregating core themes in qualitative data analysis.

Back Files: Research Material held by the R.D that can potentially be added to the repository once rights clearance and full text is obtained.

BePress: Berkeley Electronic Press public repository publishing platform.

CAIRSS: Caul Australian Institutional Repository Support Service.

CMS: Content Management System.

CODA: Institutional repository shared by NZ polytechnics and technology institutes.

Coordinator: Institutional repository coordinator responsible for liaising with depositors and the R.D, and LTS coordinators. Also called e-resources coordinator, to encompass the variety of other discovery service roles involved beyond the repository.

CC, Creative Commons licence: Rights waiver form between licensor and licensee, abridging the all rights reserved agreement for allow creative re-use of information without profit.

Crosswalked platform: Technology used in repository software such as DSpace to map identifying schemas for records across an institution's collection.

Dark Archive: Storage space in the repository that is not publicly accessible. Dark archives store items that either the author or the coordinator has chosen to preserve, but not publish. Examples are datasets, embargoed items, and items lacking full text such as abstracts which were only provided for PBRF purposes.

DC: Dublin Core Metadata Standard, DCMI: Metadata Initiative.

DIA: Department of Internal Affairs.

DNZ: Most current name for the National Library group KRIS.

DOI: Digital Object Identifier.

DSpace: Open source repository software package for creating customised content and document management systems in the publication of scholarly material.

EDUCAUSE: Non-profit association for promoting higher education with information technology.

ETD: Electronic Thesis and Dissertations.

FTE: Full Time Equivalent.

Grey Material: Non-academic articles such as magazines, news items, or conference posters. These are either published as part of the institution's cultural heritage, or ignored because they would dilute the research value of the repository collection.

Harvesting: Collecting metadata records on a selective or domain basis.

Information Platform: The online search function provided by an institution and its repository that includes its API and CMS functions to enable users to find, identify, and retrieve resources.

IR: Institutional Repository.

KRIS: Kiwi Research Information Service, now called DNZ.

LAMP: Platform bundle of Linux, Apache, MySQL and Perl/PHP/Python.

LCONZ: Library Consortium of New Zealand, Member repositories are AUT, Waikato, Otago, Unitec, and formerly, VUW.

LTS: Library Technical Service.

Marsden fields/codes: Marsden Research Classification system.

MET/MAT: Metadata Extraction/Analysis Tool.

MAG2012: The updated metadata aggregation guidelines published by DNZ for harvesters and contributors.

Metadata quality: The completeness of DC fields describing a record, and the proportion of a collection's average completeness.

NZ: New Zealand.

NRDS: National Resource Discovery Service.

OAKList: Queensland University of Technology publishing agreements database.

OAI-PMH: Open Access Initiative Protocol for Metadata Harvesting.

OECD codes: Research Classification Hierarchy for the OECD countries.

OPAC: Online Public Access Catalogue.

Open Coding: The line by line conceptualisation of transcript notes at the first level of abstraction.

Open Access: The provision of unrestricted digital access to scholarly material such as books, conference material, and journal articles.

Green Access: The provision of post prints to an Open Access archive.

Gold Access: The provision of post prints to an Open Access archive for a institutional fee, lower than corporate publishing.

Oracle: International Object Relational Database Management System.

PBRF: Performance Based Research Fund.

Public Indexers: Google, Baidu, Bing.

R.D: Research Department.

RDA: Resource Description and Access.

RDF: Resource Description Framework.

Research.digital.nz: KRIS's information platform.

RMS: Research Management System.

Shared Search Infrastructure: The service group at the National Library behind KRIS and DNZ.

SHERPA/RoMEO: UK publishing agreements directory.

SKOS: Simple Knowledge Organisation System.

SWORD: Simple Web service Offering Repository Deposit.

Symplectic: Integrated Research Management System.

TEC: Tertiary Education Commission.

XMLUI: Modular Interface layer for customising DSpace user interface.

WCT: Web Curation Tool.

WorldCat: OCLC international catalogue.

Endnotes

1. Finch, J., "Accessibility, sustainability, excellence: how to expand access to research publications", (2012), Report of the Working Group on Expanding Access to Published Research Findings. p.6
2. Ibid, p.135
3. Mustatea, Nicoleta, "To what extent is material in Institutional Repository representative of an institution's research output?" (December 2008) VUW INFO 580
4. Reid, Stephanie Frances., "An investigation into the motivating factors behind the use or non- use of institutional repositories by selected university academics", (2008) VUW INFO 580
5. Chawner, Brenda., Cullen, Rowena., "Institutional Repositories, Open Access, and Scholarly Communication: A Study of Conflicting Paradigms." *Journal of Academic Librarianship*, Vol. 37 No. 6 (2011), p.466.

6. Dickinson, Lisa., "Subject access in New Zealand Institutional Repositories: what is the best method with consideration to the self-archiving author", (2008) VUW INFO 580
7. Revell, J. T.. Perspectives of subject librarians on institutional repositories as an information resource, (2008) VUW INFO580
8. Hayes, Leonie, Morgan, Teula, Ruthven, Tom (eds), Research Repository Case Studies, EDUCAUSE Australia 2009. pp.64-109
9. Warner, Simeon., "E-prints and the Open Archives Initiative", *Library Hi Tech*, (2003), Vol. 21, No. 2, pp. 156
10. Hanlon, Ann, Ramirez, Marisa, Asking for Permission: A Survey of Copyright Workflows for Institutional Repositories, *portal: Libraries and the Academy*, Vol. 11, No. 2 (2011), p.685
11. Hanlon, Ann, Ramirez, Marisa, Asking for Permission: A Survey of Copyright Workflows for Institutional Repositories, *portal: Libraries and the Academy*, Vol. 11, No. 2 (2011), p.693
12. Payne, Nigel., Thelwall, Mike., "Do academic link types change over time?", *Journal of Documentation*, Vol. 64 No. 5, (2008), pp. 707-720
13. Strauss, A.L., & Corbin, J., Basics of qualitative research: Grounded theory procedures and techniques. (1990) Newbury Park, CA: Sage
14. Lincoln, Y.S., & Guba, E.G., "Naturalistic inquiry", (1985), Beverly Hills, CA: Sage, p.300
15. Cresswell, John W., "Research Design, Qualitative and Quantitative approaches", (1994), Thousand Oaks California, Sage. p.159
16. Mishler, E.G. (1986), Research Interviewing: Context and Narrative (Cambridge, Mass.: Harvard University Press), p.77
17. Strauss, Anselm L, Corbin, Juliet M., Basics of qualitative research: Techniques and procedures for developing grounded theory. (1998) Thousand Oaks: Sage Publications. p.125
18. Dickinson, Lisa., "Subject access in New Zealand Institutional Repositories: what is the best method with consideration to the self-archiving author", (2008) VUW INFO 580, p.41
19. HERDC: Morgan, Teula, et al., ARROW HERDC Working Group interim report May 2008.
20. Lewis, Stuart., de Castro, Pablo., Jones, Richards., "SWORD: Facilitating Deposit Scenarios" *D-Lib Magazine*, (2012), Vol. 18, No. 1/2.
21. Chawner, Brenda., Cullen, Rowena., "Institutional repositories: assessing their value to the academic community", *Performance Measurement and Metrics*, Vol. 11 No. 2, (2010), p. 145

Bibliography

Birrell, D., et al. "The DiSCmap project: Digitisation of special collections: mapping, assessment, prioritisation", *New Library World*, Vol. 112 Iss: 1, (2011), pp. 19 – 44.

Brown, A., et al, "the Library Consortium of New Zealand's Shared IRR Infrastructure" (2012)
https://docs.google.com/viewer?ae-v&q=cachet:uLEGO54bgQjwww.lconz.ac.nz/LCoNZ%2520shared%2520IRR%2520Infrastructure%2520paper%2520presented%2520at%2520the%2520Open%2520Repositories%2520Conference%25202012.pdf&hl=en&gl=nz&pid=b&srcid=ADGEE5IH4r_KeAhtoxHMyAUd_M0jCaERHICVjeWREWnp_wEp5pkpIRgz5sMDAASmkSYaoXdyanicwElyAB2Ez_shmWGh4L6Kz1w_qY4FinipRQjDaX_Gus8khh-wPbinUxpjDHK&sig=AHIEtbQxjDu5PKyMqCrlXoOzo0LsM1_Ow

Chawner, B., Cullen, R., "Institutional repositories: assessing their value to the academic community", *Performance Measurement and Metrics*, Vol. 11 No. 2, (2010), pp. 131-147.

Chawner, B., Cullen, R., "Institutional Repositories, Open Access, and Scholarly Communication: A Study of Conflicting Paradigms." *Journal of Academic Librarianship*, Vol. 37 No. 6 (2011), pp. 460-470.

Chen, X., "Google Scholar's Dramatic Coverage Improvement Five Years after Debut," *Serials Review*, (2010), Vol 36, pp. 221–226.

Craven, T., "Web authoring tools and meta tagging of page descriptions and keywords", *Online Information Review*, (2005), Vol. 29 Iss: 2 pp. 129 – 138.

Cresswell, J., "Research Design, Qualitative and Quantitative approaches", (1994), Thousand Oaks California, Sage.

Dickinson, L., "Subject access in New Zealand Institutional Repositories: what is the best method with consideration to the self-archiving author", (June 2008) VUW INFO 580

Dunsire, G, "Collecting metadata from institutional repositories OCLC Systems & Services", *International digital library perspectives* Vol. 24 No. 1, (2008), pp. 51-58

Ding, S, et al, "Batch Query Processing for Web Search Engines", *WSDM'11*, (February 9–12, 2011), Hong Kong, China.

Gross, T., Taylor, A.G., "What Have We Got to Lose? The Effect of Controlled Vocabulary on Keyword Searching Results", *College & Research Libraries*, (May 2005), pp.212-230

Hanlon, A, Ramirez, M., "Asking for Permission: A Survey of Copyright Workflows for Institutional Repositories", *portal: Libraries and the Academy*, Vol. 11, No. 2 (2011), pp. 683–702.

Hayes, L, Morgan, T, Ruthven, T (eds), "Research Repository Case Studies", EDUCAUSE Australia 2009.
<http://hdl.handle.net/2292/3368>

Hayes, L, et al. "Summary of the Discussion by IR Administrators – Held 2/5/11 at Wellington NZ", (2011)

Herrera, G, "Google Scholar Users and User Behaviors: An Exploratory Study" *College & Research Libraries* (July 2011), pp.316-331

- Huang, J., Efthimiadis, E.N., "Analyzing and Evaluating Query Reformulation Strategies in Web Search Logs", *CIKM'09*, November 2–6, (2009), pp. 1-10
- Jordan, M., "The CARL metadata harvester and search service", *Library Hi Tech*, Vol. 24 No. 2, 2006 pp. 197-210
- Kim, J., "Motivating and impeding factors affecting faculty contribution to institutional repositories", *Journal of Digital Information*, (2007), Vol. 8 No. 2.
- Lewis, S., Hayes, L., "Pimp My RMS" The University of Auckland Library, Auckland, New Zealand, Sydney, Australia 3-6 April (2011) CCA-EDUCAUSE Australasia Conference
- Lewis, S., de Castro, P., Jones, R., "SWORD: Facilitating Deposit Scenarios" *D-Lib Magazine*, (2012), Vol. 18, No. 1/2.
- Lincoln, Y.S., & Guba, E.G., "Naturalistic inquiry", (1985), Beverly Hills, CA: Sage
- Mikhail, Y., et al., "DAR: Institutional Repository Integration in Action", *TPDL 2011, LNCS 6966*, (2011), pp. 348–359, © Springer-Verlag Berlin Heidelberg
- Mishler, E.G., "*Research Interviewing: Context and Narrative*", (1986) Cambridge, Mass.: Harvard University Press.
- Morgan, T., et al., "ARROW HERDC Working Group interim report" May 2008.
<http://researchbank.swinburne.edu.au/vital/access/manager/Repository/swin:10361>
- Mustatea, N., "To what extent is material in Institutional Repository representative of an institution's research output?" (December 2008) VUW INFO 580
- Nichols, D.M., Paynter, G.W., et al, "Experiences in deploying metadata analysis tools for institutional repositories." *Cataloging & Classification Quarterly*, 47 (3/4), (2009), pp. 229-248.
- Nichols, D. M., et al., , "Metadatapedia: A proposal for aggregating metadata on data archiving", *iConference*, (2012), <http://doi.acm.org/10.1145/2132176.2132224>
- Payne, N., Thelwall, M., "Do academic link types change over time?", *Journal of Documentation*, Vol. 64 No. 5, (2008), pp. 707-720
- Pickard, A. J. "Research methods in information". (2007), London: Facet
- Qin, J., Zhang, F., "Managing Digital Repositories Through an Ontology-Based Design", *ICADL 2004, LNCS 3334*, (2004), pp. 300–309,
- Reese, T., "Automated Metadata Harvesting: Low-Barrier MARC Record Generation from OAI-PMH Repository Stores Using MarcEdit", *LRTS*, 53 no2, (April 2009)
- Reid, S., "An investigation into the motivating factors behind the use or non use of institutional repositories by selected university academics", (2008) VUW INFO 580

- Revell, J. T., "Perspectives of subject librarians on institutional repositories as an information resource", (2008) VUW INFO580
- Schroeder, R., "Pointing Users Toward Citation Searching: Using Google Scholar and Web of Science," *portal: Libraries and the Academy*, Volume 7, Number 2, (April 2007), pp. 243-248
- Shepherd-Walwyn, E., "Usage and Impact Factor Correlations in Electronic Journals", (June 2009) VUW INFO 580
- Strauss, A.L., Corbin, J.M., "Basics of qualitative research: Techniques and procedures for developing grounded theory", (1998) Thousand Oaks: Sage Publications.
- Pu, H., et al., "Subject Categorization of Query Terms for Exploring Web Users' Search Interests", *JASIST*, 53(8), (2002), pp. 617-630
- De Virgilio, R., et al., "A scalable and extensible framework for query answering over RDF" (January 2011), Springer Science+Business Media, LLC 2011
- Warner, S., "E-prints and the Open Archives Initiative", *Library Hi Tech*, (2003), Vol. 21, No. 2, pp. 151-158
- Wang, F.F., "An empirical study on the search engine optimization technique and its outcomes", Artificial Intelligence, Management Science and Electronic Commerce (AIMSEC), 2011 2nd International Conference
- Zavalina, O.L., "Contextual Metadata in Digital Aggregations: Application of Collection-Level Subject Metadata and Its Role in User Interactions and Information Retrieval", *Journal of Library Metadata*, 11:3-4, (2011), pp.104-128.
- Zuber, P. "A study of institutional repository holdings by academic discipline", *D-LibMagazine*, Vol. 14 Nos 11/12. (2008),

Websites

- CAIRSS 2011 Repository Managers Survey Report: <http://cairss.caul.edu.au/cairss/repository-manager-tools/repository-managers-survey/cairss-2011-repository-managers-survey/>
- KRIS/DNZ: <http://research.digitalnz.org/>
- NRDS 1.0 (2007): <http://www.natlib.govt.nz/downloads/NRDS-Project-Metadata-Guidelines.pdf>
- MAG 2012:
- <http://www.natlib.govt.nz/downloads/NZResearch-metadata-aggregator-guidelines.pdf>
- <http://www.natlib.govt.nz/downloads/NZResearch-metadata-guidelines-for-contributors.pdf>
- <http://www.natlib.govt.nz/downloads/NZResearch-additional-metadata-guidelines.pdf>

NZResearch Metadata Dashboard: <http://metadata.digitalnz.org/nzresearch>

List of NZ IR's sampled

[University of Canterbury Research Repository: http://ir.canterbury.ac.nz/](http://ir.canterbury.ac.nz/)

[ScholarlyCommons@AUT: http://aut.researchgateway.ac.nz/](http://aut.researchgateway.ac.nz/)

[Otago University Research Archive: http://otago.ourarchive.ac.nz/](http://otago.ourarchive.ac.nz/)

[Lincoln University Research Archive: http://researcharchive.lincoln.ac.nz/dspace/](http://researcharchive.lincoln.ac.nz/dspace/)

[ResearchCommons@Waikato: http://researchcommons.waikato.ac.nz/](http://researchcommons.waikato.ac.nz/)

[Victoria University of Wellington ResearchArchive: http://researcharchive.vuw.ac.nz/](http://researcharchive.vuw.ac.nz/)

[Massey Research Online: http://mro.massey.ac.nz/](http://mro.massey.ac.nz/)

[Unitec Research Bank: http://www.unitec.ac.nz/aboutus/research/research-bank/research-bank_home.cfm](http://www.unitec.ac.nz/aboutus/research/research-bank/research-bank_home.cfm)

Appendices

Contents

A. Information Sheet

B. Consent Form

C. Sample transcript page

D. Summary of the Discussion by IR Administrators – Held 2/5/11 at Wellington NZ, compiled by Leonie Hayes.

E. NZ Research advanced search no date, type tables with whole counts from searching without filters.

F. NZ metadata dashboard averages and complete DC field scores by Institution.

A. Information Sheet

INFORMATION SHEET

SCHOOL OF INFORMATION MANAGEMENT

I am a Masters student in Library and Information Studies at Victoria University of Wellington. As part of this degree I will be undertaking a research project that examines the developments in coordination and collaboration between New Zealand's academic Institutional Repositories and greater bodies such as DNZ, ANDS, and CAIRSS.

I am inviting Coordinators and Assistants to share their experiences in developing discovery systems for their Repository, and building working relationships with partner institutions. The main focus of this study will be examining the moves since 2009 towards the Shared Search Infrastructure (SSI) offered by DNZ. Major points of interest in this study will be the reforms offered in the Metadata Aggregation Guidelines published in March 2012.

I will be conducting interviews at a convenient time for the participants, taking an audio recording as well as written notes. The interviews will be one off sessions of approximately twenty minutes in duration. The recording and notes will be forwarded to the participant within a week of the interview. The participant may withdraw or alter their comments from inclusion in the project before midnight on the 12th of October, in which case, the data will be destroyed.

The study will use any information gathered by interviews to draw generalised, unattributed conclusions on the results of the usage analysis. The information will be used as verbal evidence of clear trends in the usage data, such as batch uploads of historic theses, conference material, and datasets. These trends will be used as evidence in the discussion section of the report regarding the history of New Zealand's repository development. This evidence will be then be used to establish a point of comparison with similar studies from other countries.

If you would like further information about the project, please contact me at bensemgreg@myvuw.ac.nz or on 021 130 7677. You may also contact my supervisor Professor Bob Allen at the School of Information Management at Victoria University of Wellington at bob.allen@vuw.ac.nz or 04 463 5887.

Greg Benseman

G J Benseman

STRUCTURED QUESTIONS LIST FOR INTERVIEW

1. How does your repository currently interact with other academic repositories in New Zealand?
For example, does your group share technical support for DSpace issues with other groups?

2. In what way does your repository group participate at the national level?

For instance, did you participate in the KRIS meeting on 2/5/11?

3. What is your group's position on maintaining or abandoning the use of the following subject coding systems:

- a) Marsden/ANZSRC.
- b) PBRF/OECD.
- c) Qualified Dublin Core.

4. In terms of expanding the collection, what do you feel is your repository's best success story in the face of challenges such as copyright licencing and depositor indifference?

5. a) How do you as a coordinator, see your repository developing in the future?

b) What obstacles will be need to be overcome to make your repository an embedded part of New Zealand's discovery infrastructure?

c) What obstacles will be need to be overcome to make your institution's repository an integrated access point of the Shared Search Infrastructure provided by the National Library?

B. Consent Form

SCHOOL OF INFORMATION MANAGEMENT

CONSENT TO PARTICIPATION IN RESEARCH

- ☐ I have been provided with adequate information relating the nature and objectives of this research project. I have understood that information and have been given the opportunity to seek further clarification or explanations.
- ☐ I understand that any information or opinions I provide will be kept confidential and reported only in an aggregated or non-attributable form.
- ☐ I understand that I will receive an audio and written transcript within a week of the interview, and have the opportunity to withdraw or amend my comments and/or request the surveyor's copy is destroyed by the 12th of October 2012.
- ☐ I understand that the audio recording, and written transcripts collected for the research project, will be destroyed within one year of its completion.
- ☐ I would like to receive a summary of the results of this research when it is completed.
- ☐ I agree to take part in this research.

Signed (typed is appropriate):

Name of participant

Please return signed and checked consent form to: bensemgreg@myvuw.ac.nz

Or post to

Mr Gregory Benseman
W.J. Scott Education Library
Panckhurst Block,
12 Donald Street,
Wellington 6012
New Zealand

C. Sample Transcript Page

G: So how do you feel about natural language versus the Marsden codes? Because presumably when digitising older theses back to 1928 a cataloguer would add codes if there weren't enough provided, but nowadays depositors tend to use natural language instead of struggling with how best to apply the codes.

R: We do try to ensure that each item that goes in has at least one Marsden or ANZSRC code, the later ones have two, sometimes three depending on the subject area. The more specialist things are we tend to offset the lack of a Marsden or ANZSRC code (from the depositor) with a couple of extra keywords from the main keywords field so we're covered from all directions possible.

G: How would you choose those keywords, are they from a controlled list?

R: They're not from a controlled list, except those of us who are doing the approval process will tend to put like things together, so we have an informal controlled list to make sure things like urban ecology items are kept together, or theses that are related to rural development, depending on the theme so for papers on rural development we make sure they all have that term somewhere in the keywords or the fields that can be picked up and collected or researched together.

G: A lot of the other institutions have found that depositors don't understand the Marsden codes, and the ANZSRC should be better, but those I've talked to have only adopted it because their partners were.

R: It's something that we'd like to keep, I think if we make too many changes to the closed sets or schemas that you use, it makes them less valuable over time, so we're persisting with them. Some of the records we have use Marsden, and later ones use ANZSRC.

G: I read in the CAIRSS EDUCAUSE 2009 document that Lincoln used to have material stored in the ADT, so there's links with Australia such as SKOS that would justify keeping the controlled vocabulary because there's still this existing library infrastructure that uses these terms.

R: Yes, and you'll probably find that Trove and other databases or collections will use something quite similar and will at least be able to relate or understand them. It's a bit like MARC cataloguing, when you're using a discovery tool like Summon, MARC is still in the background, because it's actually picking up those keywords from the MARC records in the catalog, but Summon is also discovering the Research Archive and picking them up for other reasons like Marsden and so on, so there is a relationship between those codes and schemas.

G: So do you think controlled vocabulary will still be effective in ten years?

R: I believe it will, the way we're using it is to fill gaps, but we're also using our chosen vocabulary to highlight particular subject areas, or particular themes or papers or courses which is really important to us.

G: I was talking to Emerson and he was saying that National Library lost their expertise in Oracle last year so research.digitalnz can't actually use subject headings for people browsing the

D. Summary of the Discussion by IR Administrators – Held 2/5/11 at Wellington NZ

- Discussion started by Emerson Vandy (NLNZ)
 - Overview of current support problems
 - Proposed move to DNZ infrastructure
 - Honest appraisal of possible problems (RE experience of Matapihi)
- Leonie led discussion
 - Asked for an overview to make sure we all agree - Amanda Curnow gave a 2 minute overview
 - Split into 4 groups - piece of flip chart paper each
 - 1) Do we still want to do this? Why? How?
 - 2) Data quality validation
 - 3) Opportunities of new infrastructure
 - 4) Governance
 - Group 1:
 - No need for special NZ research portal
 - KRIS is not important for users, can and do get to content in more effective ways
 - DNZ should continue to harvest repository data (unis + CRIs).
 - DNZ should provide a 'scholarly research' facet, allow the NLNZ to decide how to do this
 - Need to encourage high standards of metadata quality, but required standards for inclusion into DNZ mustn't be too high
 - Ongoing governance required (KRIS representative on DNZ board)
 - Group 4:
 - Governance needs to be sorted both ways

Quality of DNZ: Harvesting not good enough (some things in KRIS excluded from DNZ, duplicates from ARO), interface needs to be better

o Group 2:

Metadata standards - make sure abstracts are good

Get the NLNZ to do more of the 'heavy lifting'

Learn from Matapihi mistakes in transition to DNZ

Should we formalise and create a metadata group to manage and update the metadata guidelines

o Group 3:

Do we want it / why? Consensus = community support is more useful than technical aspects

Infrastructure opportunities: What if repository becomes just a black box, and use a UI provided by DNZ? Brandable interface powered by DNZ.

Browse by author, potential for national research identifier service

How do we decide on peer review

Community support (like CAIRSS)

● 5 actions:

o What other added value services could this provide (e.g. brandable UIs)

o NLNZ further develops DNZ to allow searches to be limited to scholarly research

o NZNL to make available better access to KRIS service statistics

o KRIS to address the support of active community, improve communication and resources

o Institutions creating scholarly research have a voice within SSI / DNZ governance (need to first understand DNZ governance structure)

● Outstanding questions:

o What value-added services will it provide? (E.g. suggestion of brandable service)

o Possibility of national (NZ) researcher identifier - possible? tractable? Cost effective?

- Use the NZ-IR list to seek further feedback on the Value added opportunities
- Quick review of metadata guidelines:
- Anything missing?
- Recommendation: Materials coming from our repositories are by de facto 'quality assured' therefore no need for the peer reviewed status

Summary by Leonie Hayes of our Discussions to inform the meeting of the KRIS Governance Board on the "Next Steps" paper circulated before the meeting.

1. There was agreement that if the SSI did not cost Institutions then they would contribute their metadata, but the group would like more discussion on added value and more information on other opportunities like author aggregation from the National Library (this is seen as very desirable). More details on the actual architecture and harvesting are needed (suggestion this could be contributed to Section 4 of the Metadata Guidelines).
2. The group had no problem with the service be expanded to incorporate other institutions.
3. That the Definition of "Research" be relaxed – ie include the contributions of any Institutions who were willing to meet the metadata guidelines by "lowering the entry barrier"
4. The discussion about "peer reviewed" items was inconclusive – we could not define how this can be determined, there was more interest from the group on how to define "Open Access" full text using a Metadata category – (note: can you please add more comments in the Metadata Guidelines document on how you would achieve this, section 3.8 or suggest another)
5. Governance – this conversation was wide ranging – some groups felt that representation via CONZUL to Digital NZ was adequate, others thought that far more information about the Governance Structures and how this would work was required before there could be consensus. However we all agreed that as long as there was representation and a voice for the "Research" component (i.e. that this could be articulated) then this would enable us to move forward.
6. The 4 groups had a lot to say on the Digital NZ Interface – they thought that it needed a lot of work and input from stake holders to meet the needs of the wide range of users, including identifying the user groups and their needs.

7. The group felt that the value in gathering together all NZ Repository coordinators was extremely important and creating and sustaining the community contributing Research metadata should be a high priority and achieved via any Institutional, Government, or Governance groups.

E. NZResearch Holdings counts

Type	Period	AUT	Lincoln	Massey	OpenPoly	Auckland	Unitec	Canterbury	Otago	Waikato	Victoria
Theses	1900-1999	0	1568	149		1558	0	510	50		
Conferences	1900-1999	0	281		4	74	7	25	0		
Journals	1900-1999	0	102	1	0	500	0	21	0		
Books	1900-1999	0	442		0	48	0	21	0		
Theses	2000-2009	583	932	328	1027	1331	353	1045	300	1227	941
Conferences	2000-2009	91	391	71	0	660	177	527	528	1138	
Journals	2000-2009	393	255	35	0	4383	84	314	60	3159	
Books	2000-2009	48	434	15	0	1207		67	0	246	
Theses	2010-2013	691	213		0	1205	0	321	950		
Conferences	2010-2013	0	158		0	558	0	138	0		
Journals	2010-2013	0	83	40	0	2924	0	74	0		
Books	2010-2013	0	0		0	733	0	0	0		
Other	2000-2009			1000				1898			
Other	2010-2013	469		1753			82	2000	287	43	1355
Total records		2,275	4,828	3,392	1,008	1,007	703	6,961	2,175	5,813	2,296

F. NZ Research Metadata Dashboard

Each table was last downloaded on 15/10/12

Overall Institutional Averages Table

NZResearch Metadata Dashboard	Number of Records	Average Metadata Quality
Lincoln University Research Archive	4,833	79.30%
Massey Research Online	3,390	63.10%
Open Polytechnic Repository	1,006	59.80%
Otago University Research Archive	2,111	60.00%
ResearchCommons@Waikato	5,819	59.90%
ResearchSpace@Auckland	1,007	76.50%
ScholarlyCommons@AUT	2,275	79.60%
UC Research Repository	6,972	67.20%
VUW ResearchArchive	2,297	79.90%
Unitec Research Bank	703	79.20%

Institutional DC field quality scores

Lincoln University Research Archive			
Field Name	Percent Complete	Count Valid	Count Invalid
handle	100.00%	4,796	0
title	100.00%	4,796	0
creator	100.00%	4,795	1
date	96.60%	4,633	163
eprints_type	99.90%	4,791	5
Field Name	Percent Complete	Count Complete	Count Empty
contributor	0.00%	0	4,796
description	100.00%	4,796	0
format	0.00%	0	4,796
language	99.80%	4,788	8
large_thumbnail_url	0.00%	0	4,796
marsden_code	37.10%	1,777	3,019
object_url	0.00%	0	4,796
publisher	99.80%	4,787	9
relation	38.90%	1,864	2,932

rights	4.70%	227	4,569
subject	100.00%	4,796	0
thesis_level	50.40%	2,418	2,378
thumbnail_url	0.00%	0	4,796
Massey Research Online			
Field Name	Percent Complete	Count Valid	Count Invalid
handle	100.00%	3,343	0
title	100.00%	3,342	1
creator	99.90%	3,340	3
date	14.50%	484	2,859
eprints_type	99.80%	3,336	7
Field Name	Percent Complete	Count Complete	Count Empty
contributor	0.40%	13	3,330
description	99.60%	3,330	13
format	3.70%	124	3,219
language	82.30%	2,752	591
large_thumbnail_url	0.00%	0	3,343
marsden_code	46.10%	1,541	1,802
object_url	0.00%	0	3,343
publisher	18.00%	602	2,741
relation	6.60%	221	3,122
rights	0.00%	0	3,343
subject	98.70%	3,301	42
thesis_level	13.00%	435	2,908
thumbnail_url	0.00%	0	3,343
Open Polytechnic Repository			
Field Name	Percent Complete	Count Valid	Count Invalid
handle	0.00%	0	1,051
title	100.00%	1,051	0
creator	99.30%	1,044	7
date	99.80%	1,049	2
eprints_type	0.00%	0	1,051
Field Name	Percent Complete	Count Complete	Count Empty
contributor	4.10%	43	1,008
description	99.50%	1,046	5
format	0.00%	0	1,051
language	0.50%	5	1,046

large_thumbnail_url	0.00%	0	1,051
marsden_code	99.40%	1,045	6
object_url	0.00%	0	1,051
publisher	0.00%	0	1,051
relation	0.00%	0	1,051
rights	0.00%	0	1,051
subject	100.00%	1,051	0
thesis_level	0.00%	0	1,051
thumbnail_url	0.00%	0	1,051
Otago University Research Archive			
Field Name	Percent Complete	Count Valid	Count Invalid
handle	100.00%	2,091	0
title	100.00%	2,091	0
creator	99.90%	2,089	2
date	0.00%	0	2,091
eprints_type	100.00%	2,091	0
Field Name	Percent Complete	Count Complete	Count Empty
contributor	48.80%	1,020	1,071
description	99.90%	2,088	3
format	63.80%	1,333	758
language	49.70%	1,039	1,052
large_thumbnail_url	0.00%	0	2,091
marsden_code	0.00%	0	2,091
object_url	0.00%	0	2,091
publisher	74.10%	1,549	542
relation	31.10%	651	1,440
rights	1.90%	39	2,052
subject	97.50%	2,038	53
thesis_level	65.70%	1,373	718
thumbnail_url	0.00%	0	2,091
ResearchCommons@Waikato			
Field Name	Percent Complete	Count Valid	Count Invalid
handle	96.10%	5,561	225
title	100.00%	5,786	0
creator	100.00%	5,786	0
date	0.00%	0	5,786
eprints_type	99.30%	5,743	43
Field Name	Percent	Count	Count

	Complete	Complete	Empty
contributor	10.90%	631	5,155
description	99.90%	5,779	7
format	65.60%	3,797	1,989
language	21.00%	1,216	4,570
large_thumbnail_url	0.00%	0	5,786
marsden_code	0.00%	0	5,786
object_url	0.00%	0	5,786
publisher	50.90%	2,947	2,839
relation	63.10%	3,648	2,138
rights	1.60%	91	5,695
subject	96.80%	5,602	184
thesis_level	0.00%	0	5,786
thumbnail_url	0.00%	0	5,786
ResearchSpace@Auckland			
Field Name	Percent Complete	Count Valid	Count Invalid
handle	99.90%	1,006	1
title	100.00%	1,007	0
creator	100.00%	1,007	0
date	82.70%	833	174
eprints_type	100.00%	1,007	0
Field Name	Percent Complete	Count Complete	Count Empty
contributor	5.00%	50	957
description	96.70%	974	33
format	0.10%	1	1,006
language	4.20%	42	965
large_thumbnail_url	0.00%	0	1,007
marsden_code	93.30%	939	68
object_url	0.00%	0	1,007
publisher	96.70%	974	33
relation	99.10%	998	9
rights	100.00%	1,007	0
subject	95.00%	957	50
thesis_level	6.80%	68	939
thumbnail_url	0.00%	0	1,007
ScholarlyCommons@AUT			
Field Name	Percent Complete	Count Valid	Count Invalid
handle	100.00%	2,258	0
title	100.00%	2,258	0
creator	98.90%	2,234	24

date	99.70%	2,251	7
eprints_type	99.50%	2,246	12
Field Name	Percent Complete	Count Complete	Count Empty
contributor	62.30%	1,406	852
description	97.20%	2,194	64
format	5.90%	132	2,126
language	49.20%	1,110	1,148
large_thumbnail_url	0.00%	0	2,258
marsden_code	0.00%	0	2,258
object_url	0.00%	0	2,258
publisher	99.70%	2,251	7
relation	39.70%	897	1,361
rights	1.30%	30	2,228
subject	74.60%	1,685	573
thesis_level	56.00%	1,264	994
thumbnail_url	0.00%	0	2,258
Te Tumu Eprints Repository			
Field Name	Percent Complete	Count Valid	Count Invalid
handle	0.00%	0	74
title	100.00%	74	0
creator	100.00%	74	0
date	100.00%	74	0
eprints_type	100.00%	74	0
Field Name	Percent Complete	Count Complete	Count Empty
contributor	18.90%	14	60
description	98.70%	73	1
format	97.30%	72	2
language	0.00%	0	74
large_thumbnail_url	0.00%	0	74
marsden_code	0.00%	0	74
object_url	85.10%	63	11
publisher	18.90%	14	60
relation	100.00%	74	0
rights	0.00%	0	74
subject	100.00%	74	0
thesis_level	0.00%	0	74
thumbnail_url	0.00%	0	74
UC Research Repository			
Field Name	Percent Complete	Count Valid	Count Invalid

handle	100.00%	6,935	2
title	100.00%	6,937	0
creator	100.00%	6,936	1
date	37.70%	2,614	4,323
eprints_type	98.00%	6,801	136
Field Name	Percent Complete	Count Complete	Count Empty
contributor	0.80%	54	6,883
description	95.60%	6,631	306
format	0.00%	0	6,937
language	99.90%	6,932	5
large_thumbnail_url	0.00%	0	6,937
marsden_code	32.60%	2,264	4,673
object_url	0.00%	0	6,937
publisher	100.00%	6,935	2
relation	55.90%	3,878	3,059
rights	3.70%	256	6,681
subject	71.10%	4,929	2,008
thesis_level	22.00%	1,528	5,409
thumbnail_url	0.00%	0	6,937
VUW ResearchArchive			
Field Name	Percent Complete	Count Valid	Count Invalid
handle	100.00%	2,285	0
title	100.00%	2,285	0
creator	100.00%	2,285	0
date	99.80%	2,280	5
eprints_type	99.90%	2,283	2
Field Name	Percent Complete	Count Complete	Count Empty
contributor	86.90%	1,985	300
description	100.00%	2,285	0
format	0.00%	0	2,285
language	99.90%	2,282	3
large_thumbnail_url	0.00%	0	2,285
marsden_code	0.00%	0	2,285
object_url	0.00%	0	2,285
publisher	93.40%	2,133	152
relation	0.00%	0	2,285
rights	0.30%	6	2,279
subject	99.80%	2,281	4
thesis_level	85.00%	1,942	343
thumbnail_url	0.00%	0	2,285

Unitec Research Bank			
Field Name	Percent Complete	Count Valid	Count Invalid
handle	100.00%	702	0
title	100.00%	702	0
creator	99.90%	701	1
date	100.00%	702	0
eprints_type	96.30%	676	26
Field Name	Percent Complete	Count Complete	Count Empty
contributor	46.40%	326	376
description	92.00%	646	56
format	0.00%	0	702
language	100.00%	702	0
large_thumbnail_url	0.00%	0	702
marsden_code	0.00%	0	702
object_url	0.00%	0	702
publisher	23.70%	166	536
relation	15.80%	111	591
rights	37.60%	264	438
subject	100.00%	702	0
thesis_level	0.00%	0	702
thumbnail_url	0.00%	0	702

Revised version 15 December 2012

Wordcount:12,279