Who visits academic institutional repositories and why? An analytical approach for a single high profile repository.

by

Anton Findlay Angelo

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

February 2013

Abstract

Institutional Repositories have been set up all over the world, and are now mainstream business for academic libraries and other organisations. The nature of the visitors or users for these repositories is not well understood, and little work has been done in analysing the data the repositories generate on their visitors. This report looks at the analytics generated by the University of Canterbury Research Repository (UCRR) through its own internal statistics and Google Analytics. There are many issues with reconciling this data, as many factors influence the accuracy of the figures, including web search engine crawlers, deep linking, and copyright trolls. This report found that there are many visitors to the UCRR, and that it is difficult, but possible to create narratives for specific items indicating how they might be used. Generalisations, however are much harder to make, and though we can see who is visiting the UCRR, we cannot really ascertain why they do. This report provides suggestions for further research on repository users, particularly at gathering qualitative data from groups identified from this quantative analytics.

Keywords

Analytics, Institutional Repository, Google, Green Open Access, Webliometrics, Academic Library

Acknowlegements

I would like to thank the following for their personal support during the writing of this report: Courtney Marriott, Amberleigh Nicholson, Maria Barsema, and Brenda Chawner. Also the staff at the University of Otago and University of Canterbury Libraries for their continuing encouragement for my move into an academic library career, especially Cate Bardwell, Gillian Elliot, Barbara Taylor, Anne Scott, Peter Lund and Joan Simpson.

Table of Contents

Figures

Tables

Introduction

Description of the Problem

The Ranking Web of Repositories (Consejo Superior de Investigaciones Cientificas, 2013) lists over 1,500 institutional repositories globally, but very little has been written about who uses them. This work is intended to address that by taking an analytical approach to one specific Institutional Repository (IR). By looking two at different sets of analytics this report intends to assess their relative usefulness in creating a profile of the users for this specific IR. Sub-problems of this research are to compare the different analytics sources to see if they correlate with each other, and to question how relevant exact numerical accuracy is to IR development, and if not, how can the analytics be used to answer the question of who visits an academic institutional repository, and why.

Scope and Delimitations

Definition of terms

Institutional Repositories (IRs)

> Websites that hold items of scholarly communication (e.g. journal articles, conference contributions, book chapters). The website may or may not be accessible publically. Three types of repositories have been identified, discipline, subject and institutional (Darby, Jones, Gilbert, & Lambert, 2009). Discipline and subject repositories are managed by groups of academics from multiple organisations (e.g Arxiv.org), where institutional repositories are managed by one, usually academic, organisation (e.g ir.canterbury.ac.nz). IRs normally only hold material generated by staff or students at that institution. A newer initiative is funder repositories that hold items with a link to a common funder (i.e. scoap3.org).

Items

Digital objects of scholarly communication. They could be traditional journal articles, conference papers, conference posters, video or audio items, maps, or even creative works. Any object that is an output of scholarly research or discussion that is capable of being digitised could be included in a repository. These objects could include a collection of files: a PDF article with raw research data in a spread sheet, or a PowerPoint presentation and a video.

Dark Archive

A repository that is not publically accessible, as the material on it is not able to be redistributed for copyright or other reasons (Romary & Armbruster, 2009).

Green open access

Describes a movement, or phenomenon for supporting the free availability of scholarly outputs in repositories funded and run by Universities and other public institutions (Suber, 2012).

University of Canterbury Research Repository

A DSpace implementation at http://ir.canterbury.ac.nz It is currently running on version 1.7.2, running on a Linux server hosted at the University of Canterbury in New Zealand.

Analytics

The collective term for a set of tools designed to inform business decisions. Though they could be understood as 'statistics', the term analytics implies that data has been mined and packaged to be easily understood by business users to make operational choices. "The key consumer is the business user, whose job, possibly in merchandising, marketing, or sales, is not directly related to analytics per se, but who typically uses analytical tools to improve the results of some business process along one or more dimensions ... data mining, analytic applications, and business intelligence systems are now better integrated with transactional systems than they once were, creating a closed loop between operations and analysis that

> allows data to be analysed and the results reflected quickly in business actions.” (Kohavi, Rothleder, & Simoudis, 2002, p. 45)

Bitstream

> Items in DSpace are a collection of files – usually around a single publication. A thesis item for example a thesis might include a PDF of the thesis text, a PDF appendix, and a spread sheet of raw data. Each of these is called a ‘bitstream’ as a nonspecific way to refer to any kind of file that could be uploaded. The University of Canterbury Research Repository (UCRR) includes PDFs, images, sound files, raw primary research data in a number of formats, and computer program binaries and scripts.

Scope

This study examines one high profile repository, the University of Canterbury’s Research Repository (UCRR). Described more fully below, it is New Zealand’s highest ranked repository with 90,000 unique visitors in the last six months, comparing it favourably with IRs internationally (Primary Research Group, 2014). This reasonable sample means that examining the analytics from various sources should provide a good view into the UCRR’s user base.

It is tempting to draw larger conclusions from analytics. It has to be remembered that any figures are drawn from:

1. A single site, so is not necessarily representative of IRs as a whole,
2. A specific time frame, so potentially skewed by particular events,
3. Sources designed to provide policy and business decisions, so they may not be entirely appropriate to answer specific research questions.

This study describes one IR, in the time frame of the second half of 2013. It can be compared to other IRs in very general terms (numbers of visitors, numbers of items held), but extrapolations of visitor behaviour cannot be made from one IR to another. Events, such as the Canterbury earthquakes (of which there is much related research published on the UCRR) or the Thomas Thistlewood incident (see “The Story of Thomas Thistlewood, p28) demonstrate the uniqueness of time and place to a set of data. The nature of analytics themselves show that they are not initially designed to answer research questions, but to provide guidance for business decisions. As useful as they are from a policy and operational perspective, this study necessarily limits itself to a descriptive mode, and any attempts to drawing any larger conclusions to IRs in general are very tentative, and carefully prescribed. As Fagan writes, "while web analytics may show what users are doing, they don't generally reveal why the user is doing it." (Fagan, In Press) The data in the analytics have been collected without specific research questions in mind, so assumptions made in their collection or subsequent processing may render certain questions problematic. More specific comment about this with reference to the analytics in question is made later in this report.

With those caveats on what analytics cannot do, what *can* analytics tell us? Using the two sources at this study's disposal, they can ascertain what items users want to find, and where they come from, as well as how long users spend looking for items on the site, and with a little creativity, even though we cannot be sure why most users look for most things, we can, on occasion create a plausible narrative on why some users are looking for a specific item. This study will critically examine the data collected by the two sets of analytics. It will then attempt to reconcile the two sets of data, and in the process examine what we can find out about the users of the UCRR.

Data

As this is not an experiment, with a set hypothesis to be falsified, there is no 'results' section. Rather this study will describe the data that the UCRR has available, with specific examples, and a description of the problems associated with that data. The discussion will then attempt to reconcile and synthesise that data, to assess its usefulness.

Literature Review

The literature on IRs has roughly followed three phases, each incrementally building on the previous one.

1. The technical ability for institutions to publish scholarly communication
2. Justifications and polemics for IRs, including case studies and advocacy advice
3. Examinations of the value of IRs and critical examinations on the impact of content held by IRs

The most recent, emergent, phase is developing as a concern for the quantative data that is now available for IRs. They have existed long enough to be able to do work on who is actively using them, and therefore what impact they have on the global scholarly communication system, as well as mechanisms for making knowledge available in general, outside the academy. This is part of a wider interest in how scholarly material is being used outside the ivory tower, evidenced by the growing interest in metrics included as part of the Altmetrics phenomenon. (which is regarded as out of scope of this report)

Phase One, "Because we can." The technical ability for institutions to publish scholarly communication.

The Internet lowered the cost of publishing, and as it was initially based at academic institutions, scholarly communications was one of the first applications for the new technology. Stefan Harnad's 'subversive proposal' for disintermediating scholarly communication is seen as the initial manifesto for what has become known as "green open access". (Harnad in Okerson & O'Donnell, 1995). It is interesting to note that the first discipline repository had preceded Harnad's manifesto by a number of years with the founding of http://xxx.lanl.gov/, which has become arXive.org (Ginsparg, 2011). This is a good example of how the practical application of the Internet preceded reflection and analysis on the way it disrupted previous business and social models.

Phase Two, "What are the benefits?" Justifications and polemics for IRs, including case studies and advocacy advice.

Development of IR software made it relatively simple for institutions to found their own repository. Internal business cases and justifications started to dominate the literature (Crow, 2002; Lynch, 2003; Stanger & McGregor, 2006). The aims of IRs in these documents were very specifically aimed at reforming the scholarly publishing or communication system, explicitly away from commercial publishers to a system that would be lower cost for publishing, and preferably free for the end reader. Lynch writes, "[i]t is clear that the institutional repository is a very powerful idea that can serve as an engine of change for our institutions of higher education, and more broadly for the scholarly enterprises that they support."(Lynch, 2003, p. 336).

By 2005 IRs were described as "... becoming well established as campus infrastructure components. They are broadly deployed in many of the countries surveyed, and essentially universally available in a few already." (van Westrienen & Lynch, 2005) They reported figures varied for adoption of IRs, from 5% of some European countries having up and running IRs, to other countries having 100% of their universities with at least one IR, but that the trend was for increased adoption.

Phase Three, "Critical Reflection." Examinations of the value of IRs and critical examinations of the impact of content held by IRs.

The most recent phase of research on IRs has become more critical, in both senses of showing potential failings in IRs, as well as looking more analytically at their goals and aims. Cullen and Chawner have written about the lack of recognition academics, who are often requested to voluntarily deposit items into their organisations, have of the purpose of the IR (Cullen & Chawner, 2009). Dorner and Revell examined the attitudes of academic subject librarians, and found that they were not promoting their own, or other open repositories, meaning that they were not encouraging deposit to or use of the resource. Simply put, "Input of documents + Access/use of documents = Successful IR" (Dorner & Revell, 2012, p. 262).

As IRs become more mature, studies on the cost benefit of the service (Burns, Lana, & Budd, 2013), as well as best practice recommendations (COAR, 2013) and case studies from policy and technical perspectives designed to improve the viability of the service (Giesecke, 2011; Singeh, Abrizah, & Karim, 2012) have become more common.

Emerging data, quantifying the IR

Dorner and Revell noted there are two operational sides to an IR: input of items by or on behalf of researchers, and access and use of those items by an audience, somewhere on the web (Dorner & Revell, 2012). Even though the provisioning of IR services are relatively recent, with a mean age of about 3.5 years (Primary Research Group, 2012), it is possible to start to quantify the impact of repositories. One approach is by surveying IR managers (Primary Research Group, 2008, 2009, 2012). Another is to assess IRs by looking at external metrics such as incoming links (Consejo Superior de Investigaciones Cientificas, 2013). Neither of these approaches is entirely satisfactory, as IRs differ in policy so much that to rank them against each other, as in the case of the Consejo Superior de Investigaciones Cientificas (CSIC), or assess them against mean statistics, as the Primary Research Group (PRI) have done, is problematic. The PRI's work is comprehensive, surveying a large number of repository owners, but does not analyse the results beyond basic statistics. The CSIC work ranks repositories based on a number of factors, such as the number of items, whether those items are available for download by the general public, and how 'visible' they are on internet search engines.

Looking at the question of who uses IRs, there has been little work done. In a very recent survey of US IRs, Fralinger and Bull found the majority had no idea who, outside the US, was looking at their sites (Fralinger & Bull, 2013). Why the IR audience statistics was not collected, or not able to be reported, was not well understood. "Since no statistically significant relationships were shown [between size or type of institution or type of software used for the IR], this could point to another variable or set of variables affecting the low rate of reporting, such as lack of knowledge or ability. This may also indicate a high rate of apathy regarding international usage of IRs. (Fralinger & Bull, 2013, p. 146) In short, Fralinger and Bull argue, IR managers in the United States either do not know how to examine the analytics for their IR, or do not care what the international use of an IR was. The reluctance to analyse the users visiting IRs seems to negate some of the justifications used to set up IRs described above, in the first phase of IR development. St. Jean et al. have done some valuable survey work looking at repository end users, specifically noting that little work has previously been done with this group (St. Jean, Rieh, Yakel, & Markey, 2011). Their conclusion from a small set of interviews is that repositories provide an important source of information for specific groups. Selecting which groups it would be fruitful to examine is discussed (?) in the conclusions of this report.

Method

Research Design

The IR being examined in this case is the University of Canterbury's Research Repository (UCRR?) - http://ir.canterbury.ac.nz ("UCRR: Home," n.d.). Established in 2007, it ranks as the most visible in New Zealand, and in the top 100 university repositories in the world in terms of impact (Consejo Superior de Investigaciones Cientificas, 2013). It is a DSpace based website designed to store undergraduate and postgraduate research, and staff research outputs ("UCRR - Library Wiki - Confluence," n.d.). It is not used as a dark archive, and all the metadata records it advertises have digital objects available to be freely downloaded[1].

Research Sample and Collection

The data for analysis will be drawn from two sources: the internal statistics package within DSpace, and Google Analytics. Rather than referring to experimental 'results', a description of the data available will be given, with examples. Both sets of analytics will be examined critically, with some of the problems associated with both noted, and then discussed later in the report.

Internal Statistics

The UCRR has some basic statistics gathering functions, and these have been augmented by in-house reporting tools. Though individual visits to/accesses to the UCRR are logged, the tools designed at the University of Canterbury obfuscate individual behaviour, aggregating information into topics like "most downloaded items", as well as download by faculty and type of item (i.e. theses or research outputs).

[1] Some items are embargoed for a period of up to 24 months, usually in orderto facilitate the author of a thesis to publish elsewhere. Very few publishers have an issue with theses being online while articles are prepared for publication, but there is still widespread concern from students and supervisors that having the material publically available will somehow jeopardise their publishing opportunities (Ramirez, Dalton, McMillan, Read, & Seamans, 2013).

These figures will be useful to see what visitors to the UCRR are interested in, by ranking their 'popularity'. This data has been continually collected since the UCRR has gone live, with seven years of logs.

Google Analytics

Google Analytics is a tool that collects visitor metrics for a website (Google, n.d.). By the website owner adding some tracking code on each page of their website, Google Analytics can gather information about each visitor such as; their geographic location, which pages they have been on before they visited the UCRR ('referring' sites), what search queries were used that lead the user to the IR, common search engines, and a wealth of other data. Once again the data is aggregated so that it is difficult to identify a specific user.

Google Analytics is by far the most popular tool for web analytics, used on over 40% of all sites, with 80% market share for analytic tools ("Usage Statistics and Market Share of Traffic Analysis Tools for Websites, January 2014," n.d.). Its prevalence means that it provides a way to compare websites on an 'apples to apples' basis, and could be useful in future research to see if visitor behaviour can be compared between IRs.

An issue with tracking items not being tracked by Google Analytics using the Dublin Core metadata schema in Google Scholar was identified by Arlitsch and O'Brien (Arlitsch & O'Brien, 2012), but this has been ameliorated in recent DSpace code (duraspace, n.d.) through the adoption of specific Google Scholar metadata and these mitigations implemented in the UCRR.

Google Analytics data has been collected for the UCRR since the 15th of June 2013. All figures used in this study are for the six month period ending the 15th of January 2014.

Privacy and anonymity

Raw logs gathered by the UCRR have the potential to directly identify repository users by matching internet identifiers or usernames with their behaviour. With any complex dataset involving personal information potential problems exist in linking, identifying or maintaining anonymity of individuals (Boyd & Crawford, 2012; Pfitzmann & Hansen, 2010). For this reason this study will not examine raw logs, rather it will examine aggregated reports, with no access made to any logs that could identify individuals.

Description of Collected Data

As this is not an experiment with specific results, this section describes the data that is available for the discussion, specifically for the UCRR.

Kate Marek writes, "Web analytics tools provide huge amounts of data..." (Marek in Farney & McHale, 2013, p. 26). In this case we have the two sets of data, one of which draws on every web transaction with the UCRR over seven years. The other is a detailed set of indicators already tuned for analysis, gathered over a six month period.

Analytics is the art of 'slicing and dicing' that data so that it is practically useful (and hopefully meaningful). This section is a description of what data is available, and issues with each data set.

Internal Statistics

DSpace holds a log of all web transactions for the UCRR in a Postgres database. A set of reports have been developed at Canterbury to aggregate these to help in making decisions on the service. Virtually all focus on the content of the UCRR, what items are most popular and which authors have the most downloads, and few on the users of the repository, other than what can be inferred from the former.

The reports held are (in the order they have been developed);

1. **Top 100 items**

   The most downloaded items from the UCRR, filterable by College and Department, and can be limited to the number of downloads from the last 12 months, or the life of the repository. Configurable for 10, 20 or 100 items.

2. **Top authors***

   Either the most downloaded, or the most prolific (by the number of unique items the author has deposited to the UCRR) authors, filterable by College and Department, and can be limited to the number of downloads from the last 12 months, or the life of the repository.

3. **Stats per author****

   Number of downloads per item given a specific author.

4. **Stats per item****

   Number of downloads, filterable by date range, for a specific item.

5. **Overview**

   Numbers of downloads from and deposits to the UCRR over the life of the repository or the last 12 months. Broken down by college, research type or PBRF related research outputs.

6. **PBRF-eligible outputs by College**

   The number of unique authors, items and downloads of research outputs broken down by college.

7. **Thesis count**

   The number of theses held in the UCRR, broken down to those included before and after 2005. Those written before 2005 have been added in a project to include all UC theses in the UCRR, which, at the time of writing, is close to completion.

8. **Top theses by Department**

   The most downloaded theses by department for the life of the repository, or only the last 12 months.

9. **Overview by year**

   Similar to "Overview" above, but limited to a specific calendar year.

10. **"Bubbling up" - Most downloaded recent work.**

    A configurable report of the most downloaded items filtered by college, and date range uploaded (filterable within the last 12, 18, 24 months)

11. **Grand list of Downloads and Uploads, month by month**

    Large list of all downloads and uploads, broken down by month, college, research output, and research type.

Note: All of the reports are restricted to specific UC Library staff apart from:

*report available to all UC staff

** report publically available

These reports have been developed internally in response to specific requests by managers. Each report represents a separate database querywhich is run 'live' ofor each request.

Four of these reports have been run to provide information for this study:

- 1, Top 100 items
    - 8, Top Theses by department
- 10, Bubbling up – most downloaded recent work (
- 11, Grand list of downloads and uploads, month by month
- selecting items added in the last 18 months)

Analysis of these reports will give a view of what the UCRR's visitors are looking at. The data for these reports is held in Appendix 1: Data, a spread sheet associated with this report available at http://dx.doi.org/10.6084/m9.figshare.930734 . Each of the reports is held as a separate worksheet. All of the charts below are derived from, and contained in the spread sheet.

Downloads

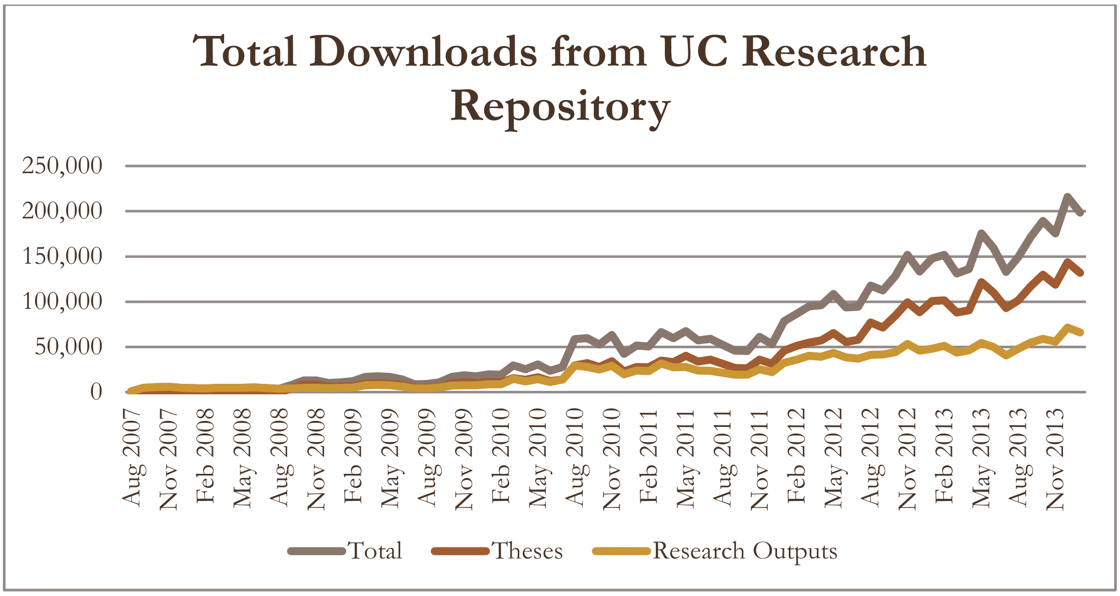
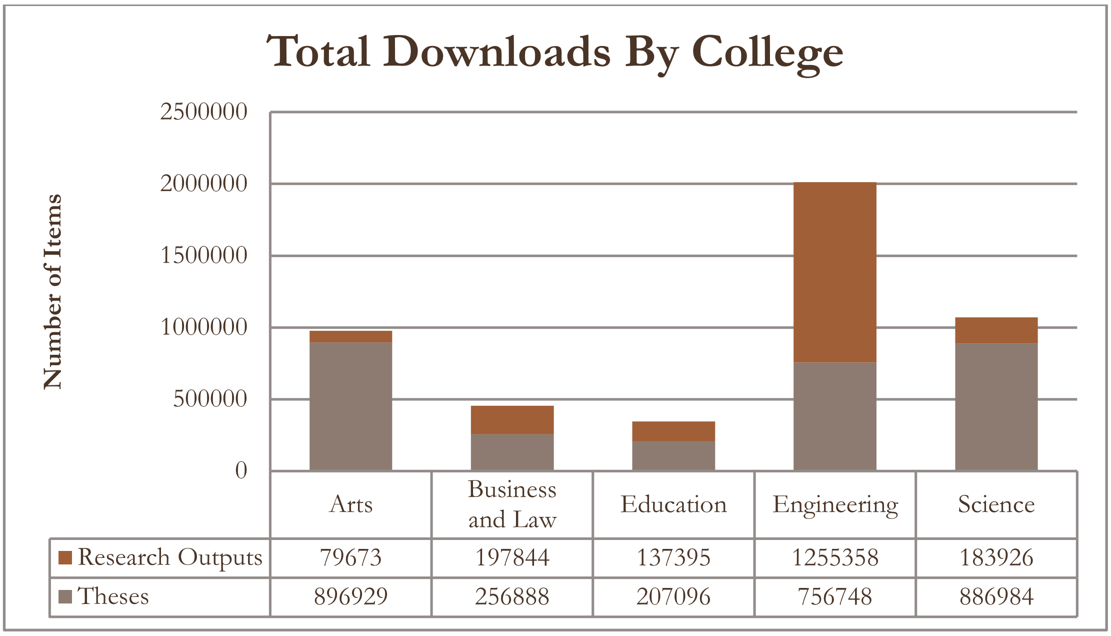


Figure 1. Total Downloads from the UCRR

Figure One shows the total downloads from the UCRR from its inception to the end of 2013. It shows a steady rise in usage shared between theses and research outputs, building over time.



	Arts	Business and Law	Education	Engineering	Science
■ Research Outputs	79673	197844	137395	1255358	183926
■ Theses	896929	256888	207096	756748	886984

Figure 2. Total Downloads by College

Thesis Date	Count
1979	1
1986	1
1993	1
1996	1
2001	1
2002	1
2004	2
2005	4
2006	13
2007	11
2008	5
2009	6
2010	3
Grand Total	50

Table 1. Count of theses in each year of the top 50 downloaded theses for the UCRR

Figure Two shows the total downloads of items broken down by College. Though this roughly reflects the number of scholars in each of the colleges, the mix of items varies considerably along disciplinary boundaries[2].

Table One shows the number of theses present for a particular year in the 50 most downloaded theses in the UCRR. This demonstrates the power of a market's 'long tail', where older material may only be accessed occasionally, but because there are many more older items than newer ones, they can form the bulk of those accessed or sold. (Anderson, 2005). In the vocabulary of commercial goods, this is facilitated by availability of older digital items which cost little to keep in inventory. A vendor, Anderson argues, can then make up the bulk of their sales from this tlong tail of older products. . The long tail goes in some way to explain why none of the top 50 theses were published in the last four years[3].

[2] The precise data for the numbers of staff working for each college is regarded as commercially sensitive, and is not included in this report.
[3] Note that all the raw data, including details of the top 50 theses, are included in the data appendix available at http://dx.doi.org/10.6084/m9.figshare.930734.

Uploads

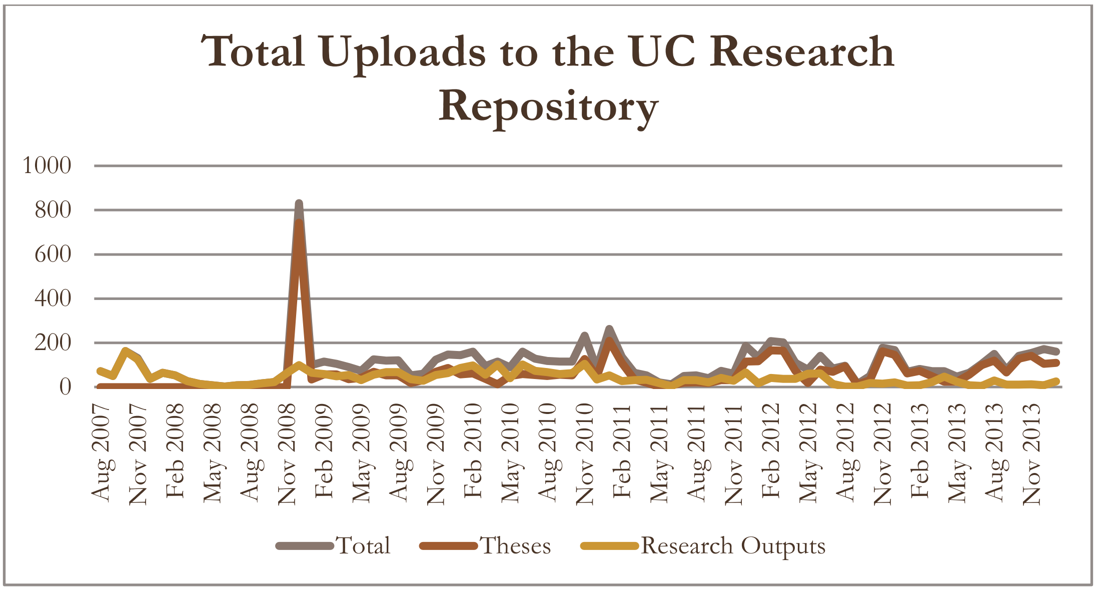
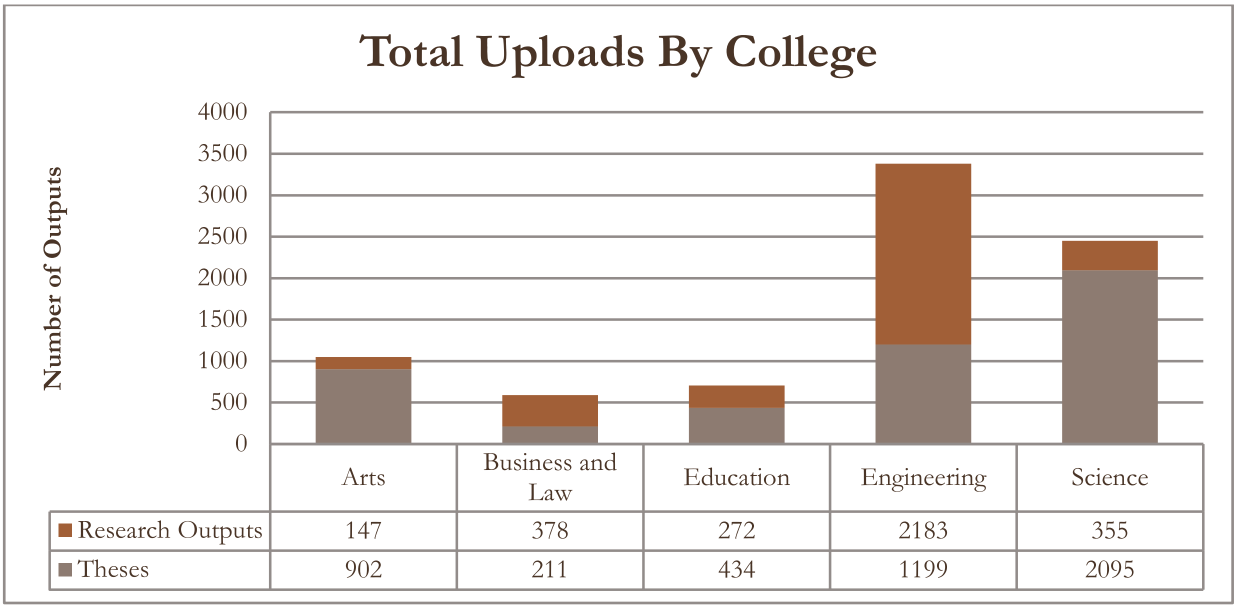


Figure 3 Total uploads to the UCRR.

Though the repository was founded in late 2007, a project to include theses retrospectively was started in 2008, and a large number were deposited late that year. Overall, the number of deposits has stayed reasonably static over the last five years.



	Arts	Business and Law	Education	Engineering	Science
■ Research Outputs	147	378	272	2183	355
■ Theses	902	211	434	1199	2095

Figure 4. Total Uploads by College

Again, as Figure Four shows, the number of uploads to the UCRR roughly reflects the number of people doing academic work in each college. The mix between theses and research outputs is similar to the downloads.

Comparing uploads and downloads

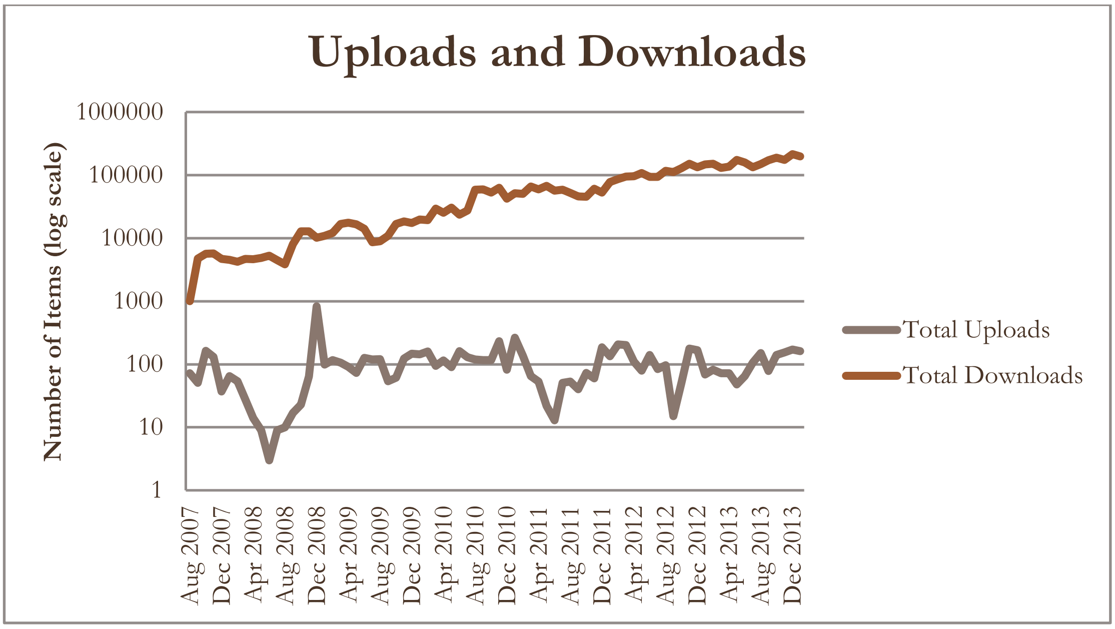


Figure 5. Uploads and Downloads for the life of the UCRR

As uploads remain static, month by month, downloads increase as shown in Figure Five. The figures have been aggregated from the uploads and downloads reports. This chart is shown in log scale in order to compare the rates of uploads and downloads, even though they are in very different scales.

Issues with the internal statistics

There are some very major problems with the data gathered automatically by DSpace, and the way they are presented here.

What constitutes a download?

Each download is intended to reflect the acquisition of an item from the repository, not just visiting a page. For every thesis download a user would normally open an item page (e.g. http://ir.canterbury.ac.nz/handle/10092/7894) and then download the bitstream (see "Definitions" above) associated with it. The bitstream may be deep linked from another site directly (e.g. the file http://ir.canterbury.ac.nz/bitstream/10092/7894/2/DataManagementLibraries_Angelo_Lund_2013.pdfis directly linked from a Google Search result). The intention of the logs, then, is to reflect how many times a bitstream has been downloaded. If an item contains more than one bitstream it could weight itself in favour of being more 'popular'.

Who, or what, is doing the downloading?

One of the main aims of the repository is to make the material on it as 'findable' as possible: to improve the chances that someone looking for material on a specific subject is presented with the option of selecting material on the UCRR. As a source of reliable information the UCRR is well indexed by major search engines, including Google, Bing and Yahoo!. In order to do that indexing, those engines acquire copies of the bitstreams in the formats they can 'read' automatically (especially PDFs) and index their contents. Some of the figures above must represent those search engines' activity, as well as indexing and harvesting carried out on the UCRR's behest like the NZresearch.org and the Online Computer Library Center's indexing programs. ("nzresearch.org.nz - Welcome to the Kiwi Research Information Service," n.d.; OCLC admin, 2013)

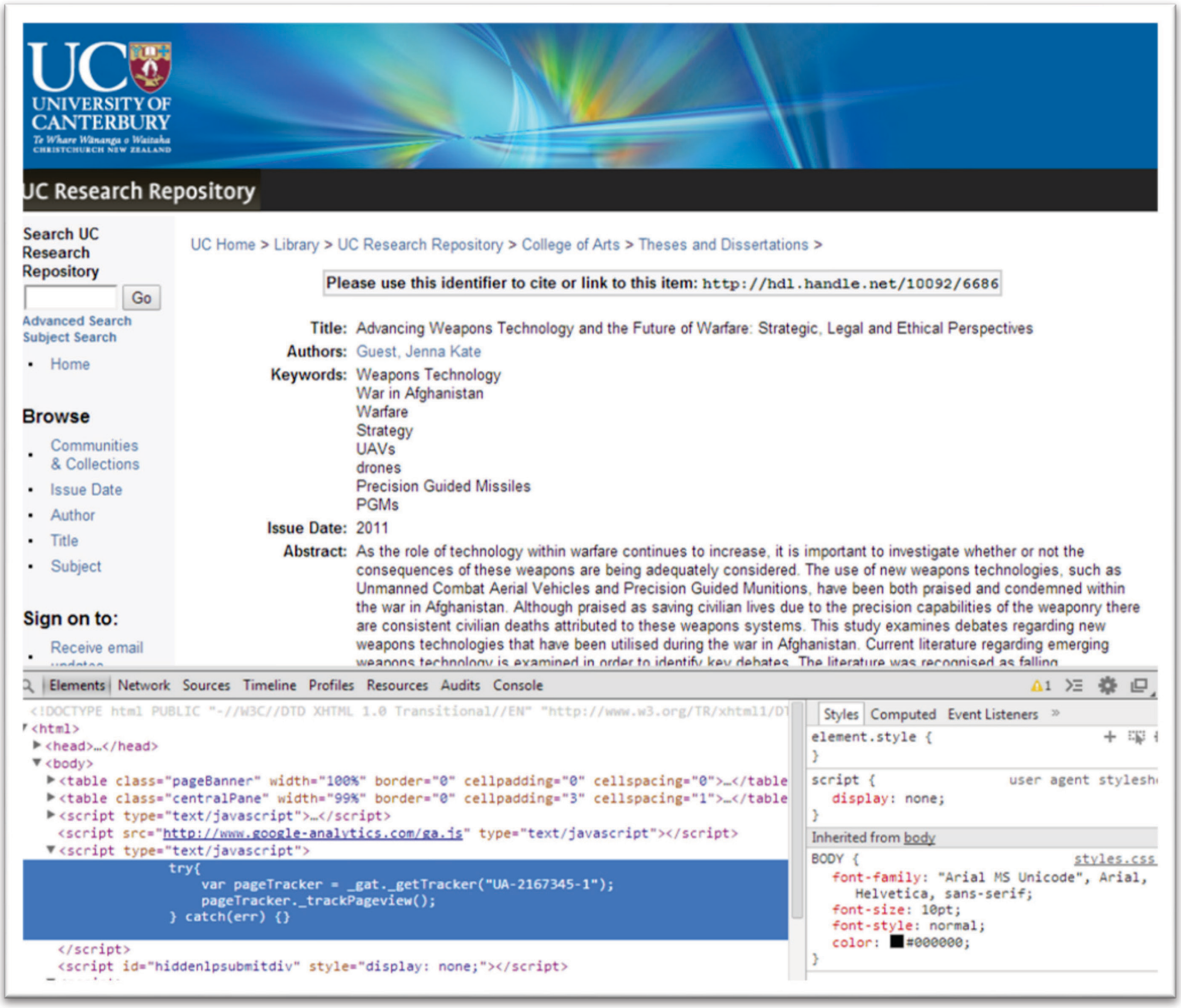
Search engine downloading tools, called spiders or robots, identify themselves by announcing themselves with a 'user agent string' as part of the hypertext transfer protocol (HTTP). Other data in that identification transaction, such as the unique internet protocol (IP) address of the computer downloading the material, or the location of any links to UCRR items in other webpages ('referrers'), are not available (Fielding, Berners-Lee, & Frystyk, 1996).

The numbers can also be inflated by other automatic 'trawling' of the UCRR bitstreams. For example organisations wanting to find copyright infringements will download all of the material in the UCRR and then look for items that contravene their publishing agreements with authors. These 'copyright troll' organisations are often paid by the number of contravening items they find, so can be very thorough. As of the time of writing the UCRR has withdrawn 21 items, only one of which was at the request of a third party copyright investigation.

These factors all artificially inflate the number of downloads reported. As the database that holds this information does not have information about specific downloaders (it does not include Internet Protocol addresses) it is impossible to filter out search engine or other harvesters.

Google Analytics

Google Analytics works by placing a small piece of code on each webpage it wants to track. As the page is accessed the code passes information about the visitor to a database at Google. Google can then draw on its own information to infer specifics about the visitor. That data is then aggregated into totals that can be broken down in many ways.



UC Home > Library > UC Research Repository > College of Arts > Theses and Dissertations >

Please use this identifier to cite or link to this item: http://hdl.handle.net/10092/6686

Title: Advancing Weapons Technology and the Future of Warfare: Strategic, Legal and Ethical Perspectives

Authors: Guest, Jenna Kate

Keywords: Weapons Technology
War in Afghanistan
Warfare
Strategy
UAVs
drones
Precision Guided Missiles
PGMs

Issue Date: 2011

Abstract: As the role of technology within warfare continues to increase, it is important to investigate whether or not the consequences of these weapons are being adequately considered. The use of new weapons technologies, such as Unmanned Combat Aerial Vehicles and Precision Guided Munitions, have been both praised and condemned within the war in Afghanistan. Although praised as saving civilian lives due to the precision capabilities of the weaponry there are consistent civilian deaths attributed to these weapons systems. This study examines debates regarding new weapons technologies that have been utilised during the war in Afghanistan. Current literature regarding emerging weapons technology is examined in order to identify key debates. The literature was recognised as falling

```
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Transitional//EN" "http://www.w3.org/TR/xhtml1/DT
<html>
 ▶<head>…</head>
 ▼<body>
   ▶<table class="pageBanner" width="100%" border="0" cellpadding="0" cellspacing="0">…</table
   ▶<table class="centralPane" width="99%" border="0" cellpadding="3" cellspacing="1">…</table
   ▶<script type="text/javascript">…</script>
    <script src="http://www.google-analytics.com/ga.js" type="text/javascript"></script>
   ▼<script type="text/javascript">
                try{
                    var pageTracker = _gat._getTracker("UA-2167345-1");
                    pageTracker._trackPageview();
                } catch(err) {}
    </script>
    <script id="hiddenlpsubmitdiv" style="display: none;"></script>
```

```
element.style {
}
script {                          user agent styleshe
   display: none;
}
Inherited from body
BODY {                                   styles.css
   font-family: "Arial MS Unicode", Arial,
      Helvetica, sans-serif;
   font-size: 10pt;
   font-style: normal;
   color: ■#000000;
}
```

Figure 6. UCRR page showing the placement of Google Analytics code.

Google Analytics was re-implemented on the UCRR in June 2013 afther having being implemented previously and failing to be updated during a DSpace upgrade. DSpace is a Java Server Pages (JSP) application, and each page is built dynamically when it is requested. The Google Analytics code was placed in a footer that appears in all pages on the site.

The results are accessible by logging into the Google Analytics website, and viewing the various reports that are available. A good detailed description of the reports available, specifically tailored to a library audience, is Farney and McHale's "Introducing Google Analytics for Libraries" (Farney & McHale, 2013). Because Google Analytics is a dynamic site new options have been made available to Google Analytics users since the publication of Farney and McHale's work, but none of those affect this report.

This report will look at a subset of information to demonstrate its capabilities; visitor numbers over time, search terms and refererring websites, geographical location, and content (pages visited).

Visitor Numbers

Metric	
Visits	118,656
Unique Visitors	101,094
Pageviews	233,630
Pages / Visit	1.97
Avg. Visit Duration	00:01:09
Bounce Rate	81.73%

Table 2 Visitor Numbers to the UCRR from Mid July 2013 - January 2014

Google Analytics gives two main metrics for visitor numbers: the total number of visitors to the site including repeat visitors (118,656) , and the unique visitor number, those that have visitied the site at least once (101,094).

Visitors went to on average almost 2 pages, stayed on the site for just over a minute, and 82% viewed one page, and left the site (bounced).This would indicate that people visit the site, and then leave immediately, without browsing the site further.

Search Tems and Referring Websites

Referrer	Visits
library.canterbury.ac.nz	2,390
canterbury.summon.serialssolutions.com	1,147
ipac.canterbury.ac.nz	1,110
scholar.google.com	884
saps.canterbury.ac.nz	523
scirus.com	471
nzresearch.org.nz	456
facebook.com	270
canterbury.ac.nz	234
scholar.google.co.uk	223

Table 3. Top 10 Referring Websites for the UCRR from Mid July 2013 - January 2014

The top three sites that referred visitors to the UCRR were the combination of University of Canterbury library sites which are comprised of the library site itself , a search engine (Summon), the library catalogue (IPAC and the School of Social and Political Science (SAPS).and). Google Scholar searches from the UK and internationally brought in almost as many visitors as the catalogue. Other notable referring sites included the now defunct scirus.com, a commercial aggregator of open research material; and nzresearch.org.nz (formerly the Kiwi Research

Information Service). Though these figures are interesting, these referrers in total represent less than 7% of all the visits to the UCRR.

Search Term	Number of visits
(not provided)	58,872
Concept of teaching	379
Thomas Thistlewood diary	209
Thesis corrosion magnesium	181
Thomas Thistlewood	178
Singapore	108
http://ir.canterbury.ac.nz/handle/10092/4612	102
The diary of Thomas Thistlewood	86
The concept of teaching	64
Thomas Thistlewood diary pdf	52

Table 4. Top 10 Organic Search Terms Leading To The UCRR

Each of the terms in Table 4 are ones that people have typed into a Google search, and then selected the link leading to the UCRR in the results. Google Analytics defines two kinds of search results: paid and organic. Since the UCRR does not have paid advertising, only organic results are presented here. The vast number of search terms are not provided by the browser.

Geographical Location

Country	Visits
New Zealand	25,733
United States	17,367
India	10,415
United Kingdom	8,134
Australia	5,089
Canada	3,172
China	2,978
Malaysia	2,893
Philippines	2,572
(not set)	2,493

Table 5. Geographical Location of UCRR Visitors

Google has a database that maps Internet Protocol networks to geographical locations, so can provide a detailed analysis of where visitors to the UCRR are coming from. Unsurprisingly, new Zealand ranks first, with the US next, followed by India and the United Kingdom.

Google Analytics can pinpoint much more accurately than just to country. For example, Figure 7 shows a graphical representation of visitors to the UCRR from India, city by city (the full data is in Appendix B: Google Analytics Reports.)

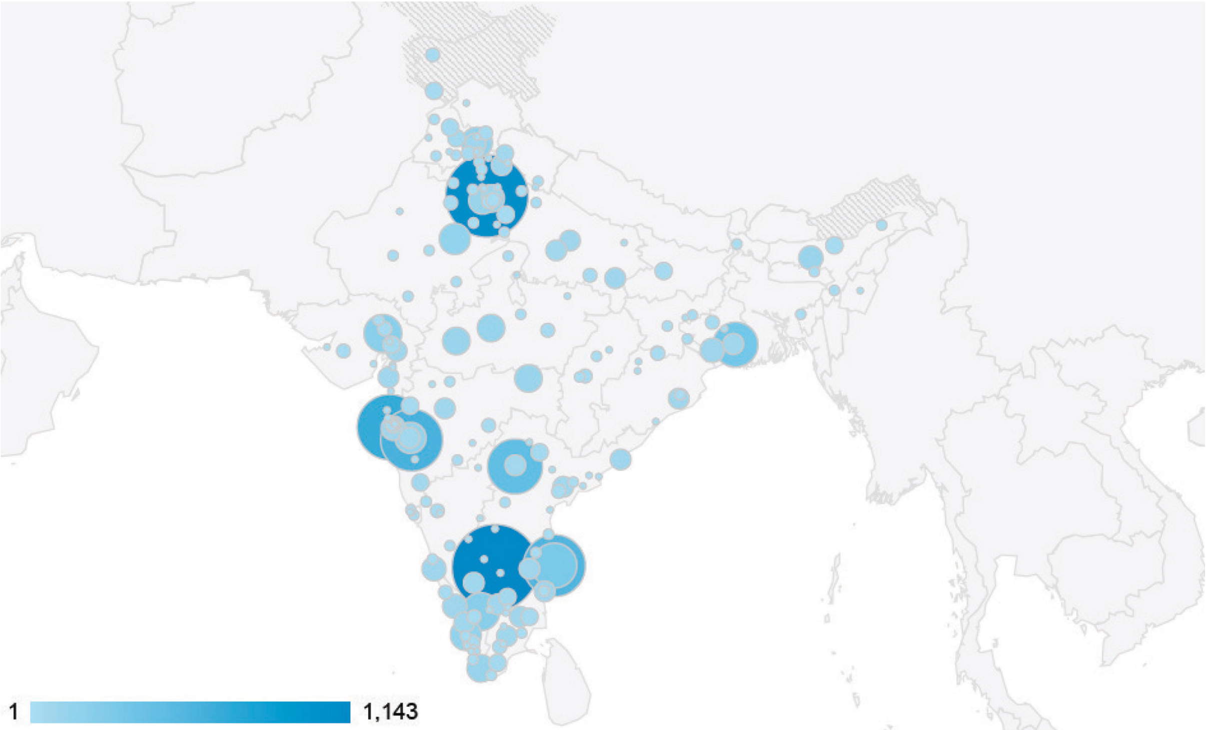


Figure 7. UCRR Visitors from Indian Cities from Google Analytics snapshot February 2014

Issues with Google Analytics

There are a number of issues with Google Analytics's figures.

As the Google Analytics code was placed on dynamically generated UCRR pages, there is a potential that some deep linking to individual bitstreams can have been missed.

Relying on Google's internal systems to aggregate data, and create inferences on things like visitor location, means that the methods for doing so are opaque. Reliability has to be taken on trust for Google's reputation for accuracy.

Discussion

The Raw and the Cooked: comparing analytics sources

Claude Lévi-Strauss used simple dichotomies within a society to unravel its sophisticated structure: raw/cooked, sacred/profane, adult/child. (Levi-Strauss, 1969). Each of these apparently simple opposites begs a question: what process takes a child to being an adult? What is the effect of declaring something sacred or profane? These questions can help us dig deep into what appears to be a simple social issue, to 'problematize' it, and let us use it as a 'tool for thinking'. In much the same way we can approach the two sources of data in this research exercise. The raw are the internal data: a mass of apparently uncomplicated numbers on visitors. The cooked are the Google Analytics data: a processed and neatly laid out web application with data finely coiffured into a form for making decisions and policy about the site the figures represent. The point Lévi-Strauss made about these dichotomies is that they both describe potential states of the same thing: a person, an object, a place. Here we have two descriptions of the same thing, the UCRR, and the data representations of the visitors to the site.

The first problem we have is reconciling the figures for the two: the numbers simply do not match. This raises questions on the validity on both sets of data. If we can somehow reconcile the figures, the next is what do they actually represent? If the figures do represent something valid, what can we actually tell about visitors to the UCRR, and the information they are acquiring from the site? Can we develop narratives about the data, or the visitors to the site, that will help us understand how it is used? Can we even point to why they might use it?

Why don't the numbers match up?

UCRR Figures, August 1st 2013 – January 31 2014		**Accesses/Downloads**
Internal Data	Total Downloads	1,100,153
Google Analytics	Visits	110,678
	Unique Visitors	94,385
	Pageviews	218,905

Table 6. Access figures for the UCRR Aug 2013-Jan 2014

As we see from Table 6, the Google Analytics and the Internal statistics for the UCRR simply do not equate to each other. Some reasons for the figures not being accurate have already been given. Deep linking to bitstreams in the UCRR could result in Google Analytics figures under-representing downloads. Not filtering for search engine spiders, robots and copyright trolls could be inflating the Internal Data results. It would be very unlikely that these figures would exactly equal each other as the methods Google use to arrive at theirs are opaque, and are subject to arbitrary change. However, these figures are not even in the same order of magnitude - can an argument be made to resolve these to each other, or to establish some kind of usefulness?

Old versus new analytics

At heart the problem can be seen as coming from two very different styles of collecting data for analytic purposes that can be explained through the history of gathering this data. In keeping with the theme of Straussian dichotomies I will describe this as the Old and the New way of collecting analytics.

Web analytics or statistics were collected from the logs of web servers, detailing how often a web page had been downloaded by a web client. Problems of what constituted a visit were worked through, and the concept of web analytics was developed[4]. Raw server logs record every file download. What constituted a web page could consist of many files: images, scripts and style sheets, all combining to create what the user saw as one 'page'. The concept of the Unique Visitor was developed in order to understand how many individuals had seen a specific site. Coming to the figure of a unique visitor required a lot of computing time, log searching and comparisons. The programmatic algorithms for doing so differed between organisations, creating a problem of comparing 'apples with oranges'. One site may have arrived at a figure quite differently from another. Standardised software - especially Open Source web statistic analysers such as the popular AWStat ("AWStats - Free log file analyzer for advanced statistics (GNU GPL).," n.d.)- ameliorated the issue, as the method for arriving at the end results was transparent.

[4] The author was personally involved in developing standards with the UK Audit Bureau of Circulations for measuring web analytics for commercial publications. Any publications/citations?

The point of transition from this Old way of looking at analytics to the New was the introduction of Google Analytics. Rather than simply presenting averages and totals of what was quite technically sophisticated data, which without experience and training was unintelligible to the average business user, Google Analytics approached the same problem from the perspective of the business user who wanted to be able to make decisions about their business. Rather than a list of details, a Google Analytics user could see, for example, if a particular geographical group were responding to specific changes to a product set advertised on a web page. As mentioned earlier, Google Analytics is able to take advantage of the understanding Google has of the internet, from knowing where specific IP networks are geographically based (not a straightforward exercise in itself), to being able to filter out its own spider and bot traffic, as well as other search engine providers' traffic.

By using the Google Analytics service different websites are able to honestly compare figures, as the algorithms behind the arrived at figures are the same. However, those algorithms, as I have mentioned, are opaque.

In this study we have Old and New style data. Though it is impossible to reconcile the difference in the figures as they are set out here, as the method used to generate one set is impossible to outline, and the other may be contaminated with programmatically generated search bot/spider results, it can be argued that both sets are useful. To do so we will use both sets of data to create narratives about how the end users of the website have found data, what data they have found, and who they are. It can be argued that the utility of the data is not in its numerical accuracy, but in its ability to confirm the importance of the IR project itself by showing how the data stored is used. When looking at specific items in the repository, we will find that the figures between the items correlate better than looking at the UCRR as a whole.

Can we tell anything about our visitors?

The data described above, as problematic as it is, tells us a lot about what users retrieve from the website, and where they access the site from. At the risk of assuming too much about the users, I want to use the data to create two narratives that can be drawn out of the data, that not only fleshes out what the users are looking for, but potentially gives a clue as to why they are doing it.

This report argues that the precise numerical accuracy of the data is questionable, but that what it can show are trends and relative rankings which, used together combine several sets of information to provide a compelling argument for the utility of the UCRR beyond a generic 'for the common good' detailed in the literature above. Creating these stories of what information needs users are having satisfied – the closest we can come to having a 'why' – provide a raison d'être that is more easily consumed than a set of raw numbers, and the assumptions implicit in their creation provide questions that can be tested, leading to better research in the future.

This report will present two narratives: one based on numerical popularity, that shows that you never know what will become popular from the 'long tail' of information in the UCRR, and another that shows that information is being provided for a small number of specific users. These two narratives, when compared, demonstrate that large and small numbers can be equally important in creating a story of how the UCRR is used.

The story of Thomas Thistlewood

Looking at the figures for organic searches in the Google Analytics data, there is one set of searches that stands out: the name "Thomas Thistlewood" is part of a search term that leads users to the UCRR. Interest in this topic started in early November 2013. Examining Google Analytics it can be seen that a spike of activity has been around one specific item: http://ir.canterbury.ac.nz/handle/10092/4674, "Thomas Thistlewood and women slaves" by Karen Rule, a PhD thesis submitted in 1994. (Rule, 1994)

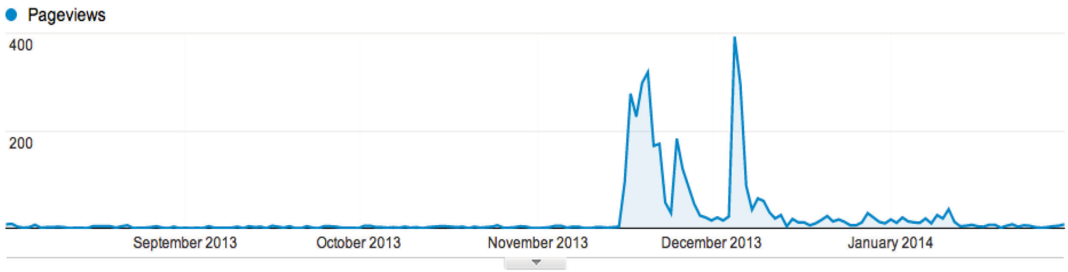


Figure 8. Pageviews for http://ir.canterbury.ac.nz/handle/10092/4674

Thomas Thistlewood was a slaver. Living in Jamaica in the 18th century his diaries detail the everyday oppressive conditions of his slaves, including the important connection between the black and white communities forged by black women who, while they were sexually exploited by white men, carried knowledge of the two communities and informed them of each other.

The answer to why Thistlewood suddenly became interesting after 10 years came when a search through news services revealed that on November 13th 2013 Martin Bashir, a US TV news anchor, criticised Alaska Governor Sarah Palin's claim that the US was a slave to Chinese financial debt.. In criticising her comparison of the US's indebtedness to 18th century US slavery, he referenced Thistlewood, and some of the appalling acts that had Thistlewood had documented in disciplining slaves he had control over. Palin struck back, and Bashir was forced to apologise, and eventually resigned his position. (Ballhaus, 2013)

A ten year old thesis in the UCRR had provided background information to a current US political stoush. Study of the analytics found that most of the visitors for that search were from the US.

The implications of the narrative we have created from this set of analytics are that the UCRR is seen as a place to get material that is background to current events. It is being ranked highly enough in popular search engines that it is being found, and acquired. It implies that the initial configuration work done by the UCRR managers and technicians in making sure that the metadata for items is accurate and being indexed correctly is working successfully.

Another lesson from this narrative is that you can never know *what* is going to be popular in the future, or *how* it is going to be made popular.

An x-ray of a lamb chop.

As well as theses and journal articles, the UCRR contains some raw data sets. Recently the MARS spectral analyser at the University of Canterbury has been producing high definition images of material (in this case, a lamb chop) as a prelude to developing a very high resolution CT scanning machine: one of orders of magnitude more detailed than current models in production (Aamir et al., 2013a).

The data published in the UCRR is very specific, and supports publications currently in press (for preprints see (Aamir et al., 2013b)). To be able to read the data in any useful way, specific sophisticated and complicated hardware and software is required, a spin off from detectors originally designed for CERN's Large Hadron Collider. The dataset is also large, about 9GB in total.

Unlike the intended end result of this research - a very high resolution CT scanner - very few people will find this raw data useful. However, the authors found it exceptionally encouraging to see that it had been accessed over 50 times, according the Internal Statistics in the UCRR. A query raised from that lead to research through Google Analytics finding that some of those downloads had occurred in the Boston area: home to other researchers who are in this field[5]. Even though the papers involved have not been formally published, the use of disciplinary and institutional preprint repositories (the UCRR and ArXiv) in tandem with the UCRR being used as a data repository lead to work being able to proceed collaboratively around the world. Though the researchers involved may well be known to each other in this rarefied field, it could also be that others who have had their interest piqued through the availability of the data now have joined the research effort.

What else can the UCRR tell us about our visitors?

The narratives above can show us something about a very small amount of the data we have available. A more prosaic reading of the analytics

What do they download the most of?

They report has argued that if we are looking for exact visitor numbers the numerical accuracy of the analytics is deeply suspect, especially for the Internal Statistics. However, when used in a way that compares items to each other, such as article rankings or finding what is becoming popular (or 'trending'), they can be much more useful. These analytics are called 'rankings', and form an important part of an on-going justification for the UCRR, and are of much interest to authors and maintainers alike.

[5] It would be possible to detail exactly which institutions (or at least, the IP networks for a specific institution) made the downloads in question, but concerns about privacy mean that this report will not go into any more detail.

Google Analytics can provide rankings for content, but the nature of the tool means it gives simple URLs, which are difficult to turn into meaningful information: rankings of the relative popularity of authors or items. The Internal Statistics, on the other hand, can give tables of the most downloaded items over time, by college and department, within a specific time frame, or uploaded within a specific time frame. Raw data for the following information is contained in Appendix 1: Data. Table 7 shows an example of the kind of data the Internal Statistics can provide, to demonstrate the usefulness of providing rankings of items, even if the exact download figures are not accurate.

Author	Submission Date	Title	Downloaded
Syed Marzuki, Sharifah Zannierah	Aug 2012	Understanding Restaurant Managers' Expectations of Halal Certification in Malaysia. Thesis for Doctor of Philosophy, University of Canterbury. Management.	3,662
Gao, Feng	Aug 2012	Pyrolysis of Waste Plastics into Fuels. Thesis for Doctor of Philosophy, University of Canterbury. Chemical and Process Engineering.	3,024
Ross, Jean C	Aug 2012	A history of poliomyelitis in New Zealand. Thesis for Master of Arts, University of Canterbury. History.	2,618
Brown, Charlotte Olivia	Sep 2012	Disaster Waste Management: a systems approach. Thesis for Doctor of Philosophy, University of Canterbury. Civil and Natural Resources Engineering.	2,170
Liu, Keqi	Oct 2012	Conscientization and the Cultivation of Conscience. Thesis for Doctor of Philosophy, University of Canterbury. Educational Studies and Human Development.	2,040
Lord, Beverley Rae, Dixon, Keith	Jan 2013	Environmental management accounting implementation in environmentally sensitive industries in Malaysia	1,962
Cochrane, Thomas	Jul 2013	La Amazonia Tierras de Bosques y Sabanas: Una guia del clima, vegetacion, paisajes y suelos de Sudamerica tropical central (Spanish edition)	1,944

Table 7. 10 Most Popular Items in the UCRR Uploaded in the Last 18 Months

It details the most popular items uploaded to the UCRR in the last 18 months. Even though the exact number of downloads may not be correct, this report assumes that the rankings of one item relative to another is correct, that any effects skewing the data is common across all items, and that any difference in the download figures is a result of the popularity of the item.

The authors and supervisors of these theses should be rightly proud of their popularity. The institution they come from is rightly interested in the spread of interest across disciplines (Engineering, Management, Philosophy…). Once again, like the Thomas Thistlewood narrative, predicting what is going to be popular is a fruitless exercise.

Author	Thesis Date	Title	Downloaded
Lloyd, Caleb Charles	2009	A Low Temperature Differential Stirling Engine for Power Generation. Thesis for Master of Engineering, University of Canterbury. Department of Electrical and Computer Engineering.	33,648
Lohmeyer, Martin	2008	The Diaoyu / Senkaku Islands Dispute. Thesis for Master of Law, University of Canterbury. Law.	27,468
Pawlowski, Ilona Paulina	2007	Sex in Women's Magazine Advertising An analysis of the degree of sexuality in women's magazine advertising across age demographics and women's responses.. Thesis for Master of Arts, University of Canterbury. Political Science and Communication.	22,255
Head, Lyndsay Fay	2006	Land, authority and the forgetting of being in early colonial Maori history. Thesis for Doctor of Philosophy, University of Canterbury. Maori and Indigenous Studies.	20,103
Ember, Adrienna	2008	Enlarged Europe, shrinking relations? the impacts of Hungary's EU membership on the development of bilateral relations between New Zealand and Hungary. Thesis for Doctor of Philosophy, University of Canterbury. Thesis for Master of Arts, University of Canterbury. European Studies.	19,623
Haslett, David Stuart	2007	Riding at the Margins: International Media and the Construction of a Generic Outlaw Biker Identity in the South Island of New Zealand, circa 1950 - 1975.. Thesis for Master of Arts, University of Canterbury. Sociology and Anthropology.	19,193
van Berkel, Haley Kathryn	2009	The Relationship Between Personality, Coping Styles and Stress, Anxiety and Depression. Thesis for Master of Science, University of Canterbury. Psychology.	18,860
Holman, Jeffrey Paparoa	2007	Best of both world: Elsdon Best and the metamorphosis of Maori spirituality. Te painga rawa o nga ao rua: Te Peehi me te putanga ke o te wairua Maori.. Thesis for Doctor of	18,255

		Philosophy, University of Canterbury. Maori and Indigenous Studies.	
Yee, Tina	2008	The Etridge influence on undue influence: attempts at fusion with duress and unconscionability. Thesis for Master of Laws, University of Canterbury. Law.	18,096
Cubrinovski, M., Ishihara, K.	2001	Correlation between penetration resistance and relative density of sandy soils. Istanbul, Turkey: 15th International Conference on Soil Mechanics and Geotechnical Engineering, 27-31 Aug 2001. 393-396. 2	16,479

Table 8. Top Downloaded Items from the UCRR to date

Table 8 shows the ranking of the most downloaded items for the UCRR to date, since the UCRR was initiated. As well as showing the mix of disciplines and topics, this also demonstrates the utility of including grey material (e.g. conference submissions) in the repository.

Summary, Conclusions and Recommendations

Summary

Through the data provided by Internal Statistics and Google Analytics, this report has attempted to discover what we can ascertain about the visitors to the UCRR.

First the type of information that the UCRR contains, and what data is collected was detailed, with a detailed set of data provided in an appendix, and summarised in a set of graphs and tables. Each of the data sources had problems identified with their data collection methods both under and over reporting the number of visitors and downloads each item received. An attempt at reconciling the figures was made, but found to be unsatisfactory.

A Straussian dichotomy was used as a model to compare and problematise the different kinds of data available to the report. The 'New' Google Analytics data, packaged, 'sliced and diced', and in a Straussian vocabulary "Cooked", was compared to the 'Old' Internal Statistics, bare, essentially unprocessed and "Raw". A short history of why the two types of data had been collected was given, and a brief discussion was made on how they could be combined, even if not numerically reconciled.

Looking at two examples of specific items it was argued that, in some specific examples, narratives about the UCRR visitors could be made. . In the Thomas Thistlewood example an event in the US media piqued interest in a topic well covered by an item in the UCRR, showing how external events can drive traffic to the UCRR. The lamb chop x-ray example shows that student created research data, of interest to a very few other researchers globally, could be confirmed to have been downloaded by other researchers in the field (please see the note about privacy as to why that could not be absolutely confirmed in this report).

Use of the numerically inaccurate data provided by the Internal Statistics could be used to provide relative rankings of data. The report argued that even if exact download figures could not be available, we can broadly tell what our visitors are interested in in general, and relative to each other, given an assumption that any data contamination was even across all items.

Conclusions

As argued in the literature review, little work has been done on the efficacy of IRs. A great amount of journal space in the Library Studies literature has gone into the arguments for initiating IRs internationally, and an entire branch of scholarly literature (Green Open Access) now relies on IRs, with governments recommending their use and research funders stipulating mandatory deposit into IRs ("About the Repository - ROARMAP," n.d.; The Working Group on Expanding Access to Published Research Findings, 2012)

It seems timely, therefore, for research to be done on the way IRs are used by their visitors. Comparing exact numbers of downloads from IRs can be done, but the results are hard to interpret in terms of the reasons why people find IRs useful. For example, surveys like the Primary Research Group's Institutional Digital Repository Benchmarks, 2014 edition, give a picture of information resources being visited and dispel the notion that they are simply being used as a write only medium (that repositories are being deposited into, but no-one is looking at them from outside) (Primary Research Group, 2014) but do not give any kind of vision of what those users are doing or why they are using the repository.

This study is really no more effective in giving a picture of a repository user. Though it details the use of the most sophisticated analytics available (Google Analytics) and has access to detailed logs of every transaction over its entire life, we are only able to build a narrative of the users for specific items in a very general way, for a very few examples. Large conclusions, like the rankings of items, and where the items are being looked at from, can be ascertained with a reasonable amount of confidence, but why those users have chosen this specific data is beyond the scope of the data available to this report.

The narratives given in the report are exceptions to the rule - it is generally hard to deduce simple narratives about why the majority of items are popular. One could speculate, but this study argues that such speculation without hard data would be pointless, and potentially misleading. Without detailed ethnographic work within each discipline, anything other than very broad assumptions about the importance of the individual work in its field would be mistaken. It could be that the work has been cited as a very poor example of its type, or used as comedic example of the sort showcased in the Ig Noble awards ("Improbable Research," n.d.).

In short, all we can tell is that the work is for some reason more popular than others: it is the work of the authors, supervisors, or further researchers to discover why.

One solid conclusion that can be made is that the UCRR is being found in Google searches (70% of the entire traffic to the UCRR comes from Google), and that the items are high enough in those search results that they are being selected and followed. The UCRR is, in Library Studies terms, 'findable'. It may seem like an obvious point to make, but it means but it means that administrative decisions such as software choice and configuration tuning have been successful, and the UCRR is in good health.

Recommendations for Future Research

Though this data does not explain why its users come to it, we can take the conclusions iterated above as a platform with which to encourage further research, and for the analytics as they are to be used in a concrete way, as part of the New Zealand academic promotion scheme.

Future Research

This report raises a number of topics that will make interesting and fruitful research in the future.

- Why is one of the most popular countries looking at the UCRR (and other repositories) India? Assumptions about educational level, language spoken and relative economic status need to be confirmed or denied.
- Who cites IR material? A bibliometric study of citations that use IR material rather than published journals may identify an inflection point in the serials crisis.
- Are IR materials more findable than published versions?

- Does making a pre-print available on an IR create a way to find that item in a subscriber based journal?
- How are institutional repositories (especially those based on different platforms) to be compared?

As well as these questions, ethnographic studies of scholars could include their methods for finding new material, their opinions on the veracity of information on an IR, and their opinions on IRs in general. Surveys of IR users could start to understand the difficult question of 'why' they have selected an IR item. Is it a simple funnel to the published version? Are they in an institution that cannot afford the final published material, like a school or a business? This work has been begun by St. Jean et al., and with some of this reports conclusions groups - such as researchers in India – could provide very fruitful results.(St. Jean et al., 2011).

Using this data

More use can be made of this data. Popularity of an item in a New Zealand (or international disciplinary) repository could be used as part of the Performance Based Research Funding process, as an indicator of research effectiveness. As well as being of interest to authors and supervisors, university marketing departments could use this information to see how people view the institution they arise from.

Bibliography

Aamir, R., Chernoglazov, A., Bateman, C. J., Butler, A. P. H., Butler, P. H., Anderson, N. G., … Hodge, K. (2013a). MARS spectral molecular imaging of lamb tissue: data collection and image analysis. *arXiv:1311.4528 [physics]*. Retrieved from http://arxiv.org/abs/1311.4528

Aamir, R., Chernoglazov, A., Bateman, C. J., Butler, A. P. H., Butler, P. H., Anderson, N. G., … Hodge, K. (2013b). MARS spectral molecular imaging of lamb tissue: data collection and image analysis. *arXiv:1311.4528 [physics]*. Retrieved from http://arxiv.org/abs/1311.4528

About the Repository - ROARMAP. (n.d.). Retrieved September 5, 2013, from http://roarmap.eprints.org/305/

Anderson, C. (2005). The Long Tail - Wired Blogs. Retrieved February 3, 2014, from http://longtail.typepad.com/the_long_tail/2005/05/the_origins_of_.html

Arlitsch, K., & O'Brien, P. S. (2012). Invisible institutional repositories: Addressing the low indexing ratios of IRs in Google Scholar. *Library Hi Tech*, *30*(1), 60–81. doi:10.1108/07378831211213210

AWStats - Free log file analyzer for advanced statistics (GNU GPL). (n.d.). Retrieved February 5, 2014, from http://awstats.sourceforge.net/

Ballhaus, R. (2013, November 22). MSNBC host's attack on Sarah Palin draws criticism. Retrieved November 25, 2013, from http://online.wsj.com/news/articles/SB10001424052702304607104579214351423986422

Boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society*, *15*(5), 662–679.

Burns, C. S., Lana, A., & Budd, J. M. (2013). Institutional Repositories: Exploration of Costs and Value. *D-Lib Magazine*, *19*(1/2). doi:10.1045/january2013-burns

COAR. (2013). *Sustainable Practices for Populating Repositories Report*. Confederation of Open Access Repositories. Retrieved from http://www.coar-repositories.org/activities/repository-content/sustainable-practices-for-populating-repositories-report/

Consejo Superior de Investigaciones Cientificas. (2013, July). July 2013 Edition. *Ranking Web of Repositories*. Retrieved September 2, 2013, from http://repositories.webometrics.info/en/node/19

Crow, R. (2002). *The Case for Institutional Repositories: A SPARC Position Paper* (p. 37). Washington DC: The Scholarly Publishing and Academic Resources Coalition.

Cullen, R., & Chawner, B. (2009). Institutional repositories and the role of academic libraries in scholarly communication. In *Asia-Pacific Conference on Library Education and Practice, Japan*.

Darby, R. M., Jones, C. M., Gilbert, L. D., & Lambert, S. C. (2009). Increasing the Productivity of Interactions Between Subject and Institutional Repositories. *New Review of Information Networking*, *14*(2), 117–135. doi:10.1080/13614570903359381

Dorner, D. G., & Revell, J. (2012). Subject librarians' perceptions of institutional repositories as an information resource. *Online Information Review*, *36*(2), 261–277. doi:10.1108/14684521211229066

duraspace. (n.d.). Google Scholar Metadata Mappings - DSpace 3.x Documentation - DuraSpace Wiki. Retrieved September 3, 2013, from https://wiki.duraspace.org/display/DSDOC3x/Google+Scholar+Metadata+Mappings

Fagan, J. C. (In Press). The Suitability of Web Analytics Key Performance Indicators in the Academic Library Environment. *The Journal of Academic Librarianship*. doi:10.1016/j.acalib.2013.06.005

Farney, T., & McHale, N. (2013). Introducing Google Analytics for Libraries. *Library Technology Reports*, *49*(4), 5–8.

Fielding, R. T., Berners-Lee, T., & Frystyk, H. (1996). Hypertext Transfer Protocol -- HTTP/1.0. Retrieved February 3, 2014, from http://tools.ietf.org/html/rfc1945

Fralinger, L., & Bull, J. (2013). Measuring the international usage of US institutional repositories. *OCLC Systems & Services*, *29*(3), 134–150. doi:10.1108/OCLC-10-2012-0039

Giesecke, J. (2011). Institutional Repositories: Keys to Success. *Journal of Library Administration*, *51*(5-6), 529–542. doi:10.1080/01930826.2011.589340

Ginsparg, P. (2011). It was twenty years ago today ... *arXiv:1108.2700*. Retrieved from http://arxiv.org/abs/1108.2700

Google. (n.d.). Google Analytics Official Website - Web Analytics & Reporting – Google Analytics. Retrieved September 2, 2013, from http://www.google.com/analytics/

Improbable Research. (n.d.). *Improbable Research*. Retrieved February 7, 2014, from http://www.improbable.com

Kohavi, R., Rothleder, N. J., & Simoudis, E. (2002). Emerging Trends in Business Analytics. *Commun. ACM*, *45*(8), 45–48. doi:10.1145/545151.545177

Lévi-Strauss, C. (1969). *The Raw and the Cooked*. London: Jonathan Cape.

Lynch, C. A. (2003). Institutional Repositories: Essential Infrastructure For Scholarship In The Digital Age. *Portal: Libraries and the Academy*, *3*(2), 327–336. doi:10.1353/pla.2003.0039

*Martin Bashir Says Someone Should Sh*t in Sarah Palin's Mouth*. (2013). Retrieved from http://www.youtube.com/watch?v=3fbJE3RUJMw&feature=youtube_gdata_player

nzresearch.org.nz - Welcome to the Kiwi Research Information Service. (n.d.). Retrieved March 15, 2012, from http://nzresearch.org.nz/index.php/index

OCLC admin. (2013, February 16). Frequently asked questions. Retrieved February 3, 2014, from http://www.oclc.org/oaister/questions.en.html

Okerson, A. S., & O'Donnell, J. J. (Eds.). (1995). *Scholarly Journals at the Crossroads: A Subversive Proposal for Electronic Publishing* (Last Edited 1998.). Washington DC: Association of Research Libraries. Retrieved from http://www.arl.org/sc/subversive/

Pfitzmann, A., & Hansen, M. (2010). A terminology for talking about privacy by data minimization: Anonymity, unlinkability, undetectability, unobservability, pseudonymity, and identity management. *Version 0.34 Aug, 10*.

Primary Research Group. (2008). *The international survey of institutional digital repositories* (2008 ed.). New York: Primary Research Group.

Primary Research Group. (2009). *The survey of higher education faculty: use of digital repositories and views on open access*. New York, N.Y: Primary Research Group.

Primary Research Group. (2012). *The survey of institutional digital repositories, 2012-13 edition*. New York: Primary Research Group. Retrieved from http://helicon.vuw.ac.nz/login?url=http://library.victoria.ac.nz/ebooks/PRG_The Survey of Institutional Digital Repositories, 2012-13 Edition.pdf

Primary Research Group. (2014). *Institutional digital repository benchmarks, 2014 edition*. New York, NY: Primary Research Group.

Ramirez, M. L., Dalton, J. T., McMillan, G., Read, M., & Seamans, N. H. (2013). Do Open Access Electronic Theses and Dissertations Diminish Publishing Opportunities in the Social Sciences and Humanities? Findings from a 2011 Survey of Academic Publishers. *College & Research Libraries*, *74*(4), 368–380.

Romary, L., & Armbruster, C. (2009). *Beyond Institutional Repositories* (SSRN Scholarly Paper No. ID 1425692). Rochester, NY: Social Science Research Network. Retrieved from http://papers.ssrn.com/abstract=1425692

Rule, K. (1994). UC Research Repository: Thomas Thistlewood and women slaves. Retrieved November 25, 2013, from http://ir.canterbury.ac.nz/handle/10092/4674

Singeh, F. W., Abrizah, A., & Karim, N. H. A. (2012). What inhibits authors to self-archive in Open Access repositories? A Malaysian case. *Information Development*, *29*(1), 24–35. doi:10.1177/0266666912450450

St. Jean, B., Rieh, S. Y., Yakel, E., & Markey, K. (2011). Unheard voices: institutional repository end-users. *College & Research Libraries*, *72*(1), 21–42.

Stanger, N., & McGregor, G. (2006). Hitting the ground running: building New Zealand's first publicly available institutional repository. Retrieved from http://otago.ourarchive.ac.nz/handle/10523/915

Suber, P. (2012). *Open Access*. The MIT Press.

The Working Group on Expanding Access to Published Research Findings. (2012). *Accessibility, sustainability, excellence: how to expand access to research publications*. London: The Research Information Network. Retrieved from http://www.researchinfonet.org/wp-content/uploads/2012/06/Finch-Group-report-FINAL-VERSION.pdf

UC Research Repository - Library Wiki - Confluence. (n.d.). Retrieved September 2, 2013, from http://wiki.canterbury.ac.nz/display/LIBRARY/UC+Research+Repository

UC Research Repository: Home. (n.d.). Retrieved September 2, 2013, from http://ir.canterbury.ac.nz/

Usage Statistics and Market Share of Traffic Analysis Tools for Websites, January 2014. (n.d.). Retrieved January 26, 2014, from http://w3techs.com/technologies/overview/traffic_analysis/all

Van Westrienen, G., & Lynch, C. A. (2005). Academic Institutional Repositories: Deployment Status in 13 Nations as of Mid 2005. *D-Lib Magazine*, *11*(09). doi:10.1045/september2005-westrienen

Appendix 1: Data

Data is available online under a CC-BY licence.

Angelo, Anton (2014): Appendix One: Data. figshare.

http://dx.doi.org/10.6084/m9.figshare.930734

Appendix 2: Google Analytics Reports

The following reports were extracted for this study from Google Analytics in February 2014.

Google Analytics

Go to this report

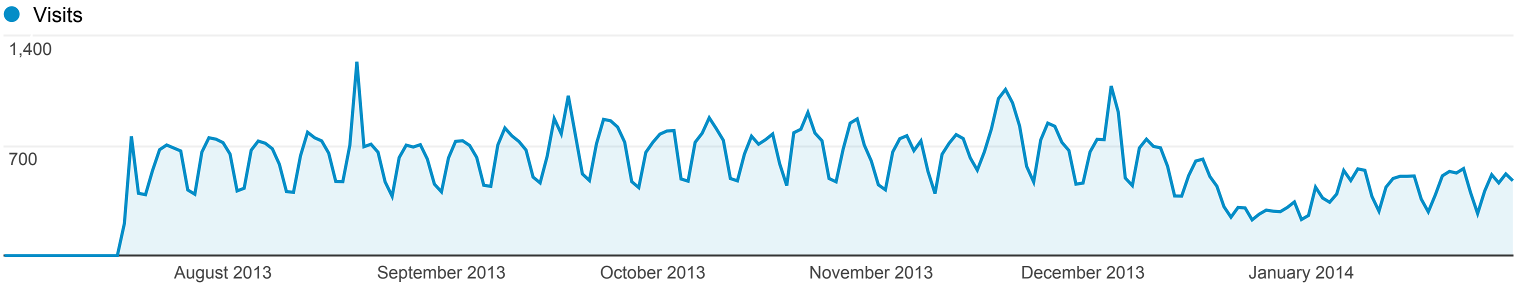
http://ir.canterbury.ac.nz - http://ir.canterbury.ac.nz
UC Research Repository

Audience Overview

Jul 1, 2013 - Jan 31, 2014

All Visits
100.00%

Overview



101,094 people visited this site

Visits
118,656

Unique Visitors
101,094

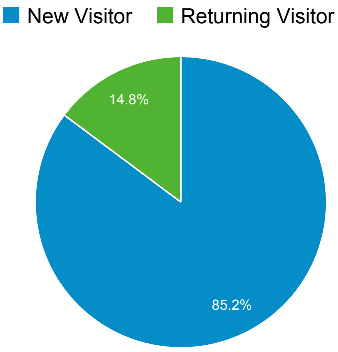
Pageviews
233,630

Pages / Visit
1.97

Avg. Visit Duration
00:01:09

Bounce Rate
81.73%

% New Visits
85.15%



	Language	Visits	% Visits
1.	en-us	83,980	70.78%
2.	en-gb	9,552	8.05%
3.	zh-cn	3,701	3.12%
4.	en	2,304	1.94%
5.	fr	1,504	1.27%
6.	es	1,424	1.20%
7.	de-de	1,312	1.11%
8.	zh-tw	927	0.78%
9.	pt-br	891	0.75%
10.	es-es	720	0.61%

Google Analytics

Go to this report

http://ir.canterbury.ac.nz - http://ir.canterbury.ac.nz
UC Research Repository

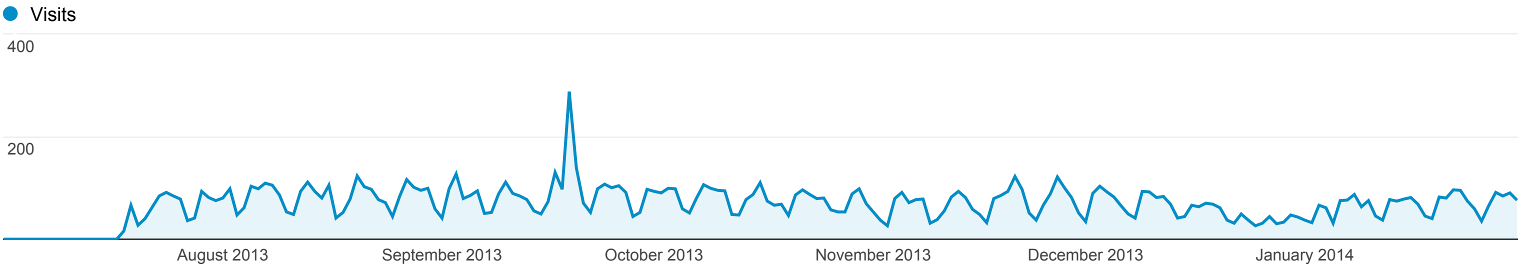
Referral Traffic

Jul 1, 2013 - Jan 31, 2014

All Visits
12.31%

Explorer

Summary



Source	Acquisition			Behavior			Conversions Goal 1: Item Description		
	Visits	% New Visits	New Visits	Bounce Rate	Pages / Visit	Avg. Visit Duration	Item Description (Goal 1 Conversion Rate)	Item Description (Goal 1 Completions)	Item Description (Goal 1 Value)
	14,604 % of Total: 12.31% (118,656)	71.39% Site Avg: 85.15% (-16.16%)	10,426 % of Total: 10.32% (101,032)	64.98% Site Avg: 81.73% (-20.49%)	4.02 Site Avg: 1.97 (104.37%)	00:02:37 Site Avg: 00:01:09 (128.75%)	61.94% Site Avg: 64.08% (-3.33%)	9,046 % of Total: 11.90% (76,029)	$0.00 % of Total: 0.00% ($0.00)
1. library.canterbury.ac.nz	**2,390**	38.41%	918	20.25%	12.21	00:06:23	50.88%	1,216	$0.00
2. canterbury.summon.serialssolutions.com	**1,147**	59.90%	687	72.45%	1.94	00:01:32	53.44%	613	$0.00
3. ipac.canterbury.ac.nz	**1,110**	60.99%	677	63.24%	3.14	00:04:34	64.50%	716	$0.00
4. scholar.google.com	**884**	89.82%	794	85.52%	1.44	00:00:46	76.02%	672	$0.00
5. saps.canterbury.ac.nz	**523**	83.17%	435	45.89%	6.88	00:03:15	67.69%	354	$0.00
6. scirus.com	**471**	85.56%	403	81.74%	1.79	00:01:08	62.21%	293	$0.00
7. nzresearch.org.nz	**456**	75.66%	345	66.45%	2.38	00:03:26	54.82%	250	$0.00
8. facebook.com	**270**	71.48%	193	63.33%	3.48	00:03:05	63.70%	172	$0.00
9. canterbury.ac.nz	**234**	76.07%	178	64.10%	3.64	00:02:09	47.44%	111	$0.00
10. scholar.google.co.uk	**223**	88.34%	197	91.03%	1.12	00:00:07	77.13%	172	$0.00

Rows 1 - 10 of 910

Google Analytics

Go to this report

http://ir.canterbury.ac.nz - http://ir.canterbury.ac.nz
UC Research Repository

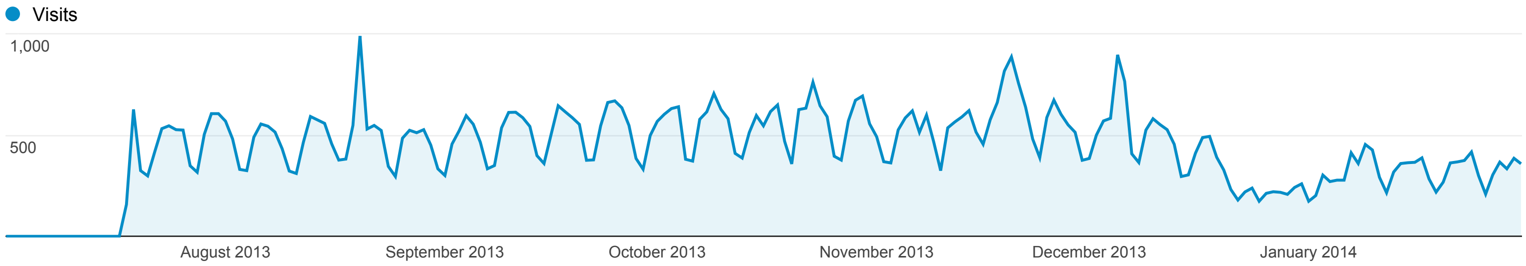
Organic Search Traffic

Jul 1, 2013 - Jan 31, 2014

All Visits
77.92%

Explorer

Summary



	Keyword	Acquisition			Behavior			Conversions	Goal 1: Item Description	
	Keyword	**Visits**	**% New Visits**	**New Visits**	**Bounce Rate**	**Pages / Visit**	**Avg. Visit Duration**	**Item Description (Goal 1 Conversion Rate)**	**Item Description (Goal 1 Completions)**	**Item Description (Goal 1 Value)**
		92,458 % of Total: 77.92% (118,656)	88.40% Site Avg: 85.15% (3.82%)	81,734 % of Total: 80.90% (101,032)	85.14% Site Avg: 81.73% (4.18%)	1.45 Site Avg: 1.97 (-26.16%)	00:00:45 Site Avg: 00:01:09 (-34.20%)	65.64% Site Avg: 64.08% (2.44%)	60,688 % of Total: 79.82% (76,029)	$0.00 % of Total: 0.00% ($0.00)
1.	(not provided)	**58,872**	89.55%	52,721	84.55%	1.47	00:00:47	71.32%	41,987	$0.00
2.	concept of teaching	**379**	88.92%	337	84.96%	1.26	00:00:53	63.59%	241	$0.00
3.	thomas thistlewood diary	**209**	95.69%	200	96.65%	1.06	00:00:11	93.78%	196	$0.00
4.	thesis corrosion magnesium	**181**	0.55%	1	90.61%	1.13	00:00:41	15.47%	28	$0.00
5.	thomas thistlewood	**178**	93.26%	166	92.13%	1.14	00:00:26	91.57%	163	$0.00
6.	singapore	**108**	93.52%	101	87.04%	1.21	00:00:37	70.37%	76	$0.00
7.	http://ir.canterbury.ac.nz/handle/10092/4612	**102**	92.16%	94	84.31%	1.45	00:00:14	70.59%	72	$0.00
8.	the diary of thomas thistlewood	**86**	96.51%	83	97.67%	1.02	00:00:12	88.37%	76	$0.00
9.	the concept of teaching	**64**	82.81%	53	85.94%	1.47	00:00:48	73.44%	47	$0.00
10.	thomas thistlewood diary pdf	**52**	84.62%	44	90.38%	1.13	00:00:14	86.54%	45	$0.00

Rows 1 - 10 of 27319

Google Analytics

Go to this report

http://ir.canterbury.ac.nz - http://ir.canterbury.ac.nz
UC Research Repository

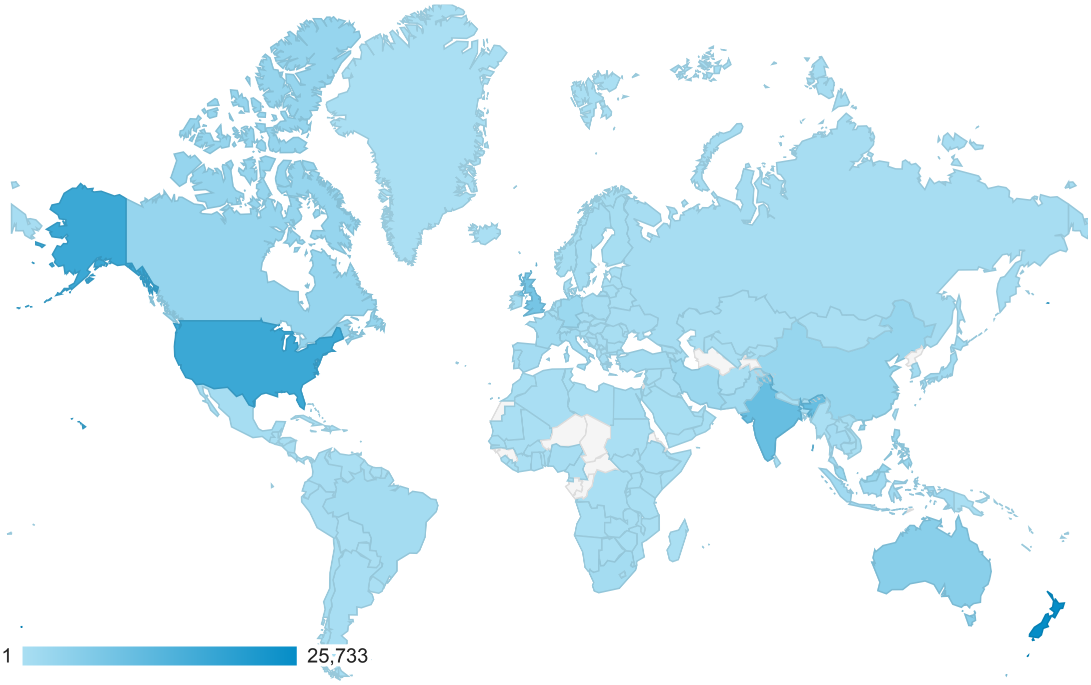
Location

Jul 1, 2013 - Jan 31, 2014

All Visits
100.00%

Map Overlay

Summary



Country / Territory	Acquisition			Behavior			Conversions Goal 1: Item Description		
	Visits	% New Visits	New Visits	Bounce Rate	Pages / Visit	Avg. Visit Duration	Item Description (Goal 1 Conversion Rate)	Item Description (Goal 1 Completions)	Item Description (Goal 1 Value)
	118,656 % of Total: 100.00% (118,656)	85.21% Site Avg: 85.15% (0.07%)	101,102 % of Total: 100.07% (101,032)	81.73% Site Avg: 81.73% (0.00%)	1.97 Site Avg: 1.97 (0.00%)	00:01:09 Site Avg: 00:01:09 (0.00%)	64.08% Site Avg: 64.08% (0.00%)	76,029 % of Total: 100.00% (76,029)	$0.00 % of Total: 0.00% ($0.00)
1. New Zealand	**25,733**	72.10%	18,553	71.28%	3.50	00:02:13	55.75%	14,347	$0.00
2. United States	**17,367**	91.92%	15,964	88.21%	1.33	00:00:36	74.49%	12,937	$0.00
3. India	**10,415**	92.18%	9,601	84.92%	1.57	00:00:59	60.30%	6,280	$0.00
4. United Kingdom	**8,134**	85.82%	6,981	85.53%	1.33	00:00:37	70.53%	5,737	$0.00
5. Australia	**5,089**	86.46%	4,400	82.81%	1.66	00:00:49	56.02%	2,851	$0.00
6. Canada	**3,172**	88.11%	2,795	86.35%	1.32	00:00:38	71.75%	2,276	$0.00
7. China	**2,978**	78.51%	2,338	80.12%	1.94	00:01:11	69.31%	2,064	$0.00
8. Malaysia	**2,893**	88.63%	2,564	84.20%	1.60	00:00:51	68.06%	1,969	$0.00
9. Philippines	**2,572**	91.45%	2,352	86.78%	1.43	00:00:48	51.75%	1,331	$0.00
10. (not set)	**2,493**	91.86%	2,290	87.20%	1.25	00:00:32	63.38%	1,580	$0.00

Rows 1 - 10 of 208

Google Analytics

Go to this report

http://ir.canterbury.ac.nz - http://ir.canterbury.ac.nz
UC Research Repository

Location

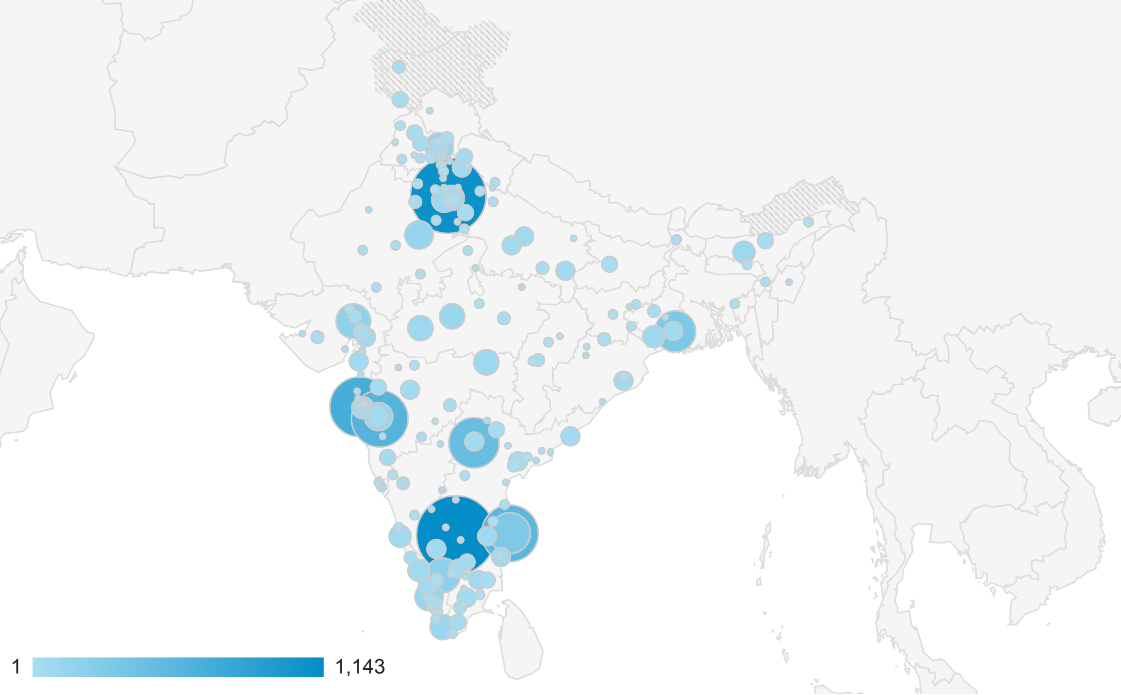
Jul 1, 2013 - Jan 31, 2014

ALL » COUNTRY / TERRITORY: India

All Visits
8.78%

Map Overlay

Summary



City	Acquisition			Behavior			Conversions	Goal 1: Item Description	
	Visits	% New Visits	New Visits	Bounce Rate	Pages / Visit	Avg. Visit Duration	Item Description (Goal 1 Conversion Rate)	Item Description (Goal 1 Completions)	Item Description (Goal 1 Value)
	10,415 % of Total: 8.78% (118,656)	92.18% Site Avg: 85.15% (8.26%)	9,601 % of Total: 9.50% (101,032)	84.92% Site Avg: 81.73% (3.90%)	1.57 Site Avg: 1.97 (-20.24%)	00:00:59 Site Avg: 00:01:09 (-14.31%)	60.30% Site Avg: 64.08% (-5.90%)	6,280 % of Total: 8.26% (76,029)	$0.00 % of Total: 0.00% ($0.00)
1. Bangalore	**1,143**	93.18%	1,065	85.83%	1.50	00:00:49	60.80%	695	$0.00
2. New Delhi	**1,085**	90.69%	984	84.42%	1.55	00:01:14	65.16%	707	$0.00
3. Chennai	**860**	90.58%	779	83.26%	1.73	00:01:05	58.14%	500	$0.00
4. Mumbai	**683**	94.73%	647	88.29%	1.32	00:00:37	61.20%	418	$0.00
5. Pune	**596**	91.28%	544	86.91%	1.37	00:00:43	62.58%	373	$0.00
6. Hyderabad	**460**	91.09%	419	84.13%	1.65	00:01:22	63.91%	294	$0.00
7. Kolkata	**323**	93.19%	301	87.00%	1.42	00:00:38	62.54%	202	$0.00
8. Coimbatore	**212**	94.81%	201	85.38%	1.54	00:00:43	56.13%	119	$0.00
9. Ahmedabad	**199**	95.48%	190	85.93%	1.45	00:00:52	56.78%	113	$0.00
10. Pimpri Chinchwad	**184**	93.48%	172	87.50%	1.59	00:00:57	52.17%	96	$0.00

Rows 1 - 10 of 192