Resource Discovery @ The University of Oxford

2015

Analysis & Recommendations by Christine Madsen & Megan Hurst, Athenaeum21 Consulting
Research by

Christine Madsen
Iain Emsley
Alfie Abdul-Rahman
Min Chen
Megan Hurst
Ray Stacey
Saiful Khan
Simon McLeish
Masha Garibyan

Contents

11 February 2016. This version for public distribution

Executive Summary & Recommendations

This project has conducted 113 interviews, 18 site visits, and 3 literature reviews in order to discover requirements of users at Oxford and understand the broader landscape of resource discovery. Through analysis of the data across all of these areas, a significant and nuanced understanding of current and future trends in resource discovery has emerged. Based on this, the following recommendations have been made.

Areas for Investment, Part 1: Mapping the Landscape of Things

- **Visualizing the scope of the collections at Oxford.** Using collection-level metadata, provide an interactive diagram that represents the range of collections at Oxford.
- **Cross collection search**. Taking existing metadata from across the collections, use a Lucene-based technology such as Elasticsearch to index and expose the existing item-level metadata.

Areas for Investment, Part 2: Mapping the Landscape of People

- **Create a directory of expertise at Oxford.** The current researcher collaboration tool project, run by Research Services, aims to provide a directory of research and expertise. Ideally, this project should be built with an open and flexible framework in order to further enable:
- **Visualization of the network of experts and research**. A graph of the professional networks at Oxford would facilitate discovery and navigation within and between fields.
- **Connect people, resources, and events.** Providing a reliable source for upcoming talks by division or subject area would be heavily used and well-received.

Areas for Investment, Part 3: Supporting Researchers' Established Practices

- **Getting existing metadata out to the places where many researchers work**. Exposing metadata for indexing by Google and Google Scholar would undoubtedly assist those who start their searches on the open web.
- **Investigate methods to facilitate citation-chaining**, which is ubiquitous across all disciplines.

Next Steps

These activities should be guided and supported by **establishing an interdisciplinary research group** to take forward the recommendations of the report.

This research group should ensure **investment in the analytics and data infrastructure to support evidence based decision making across the collections**.

The Academic Services and University Collections (ASUC) should investigate the creation of a **'Collections @ Oxford' portal.**

Activities will be integrated with other potential or actual projects, notably the researcher collaboration tool project, the Oxford Linked Open Data project, and a platform for digitized content throughout the University.

Aims and Objectives

The University of Oxford aims to lead the world in research and education.[1] This is driven by its consumption and production of priceless intellectual assets including publications and data, teaching resources, library resources, archives and museum collections. However, though some collections, resources and expertise are catalogued and listed in great detail, they can be hard to find, take diverse forms and are often not suited to discovery by potential users. And the catalogued collections only represent a part of the overall holdings of the University's museums and libraries. This means that Oxford researchers and students are not benefiting as fully as they should from the wealth of knowledge that has been collected, created, purchased or licensed on their behalf. Thus the University's riches are hidden from view, underexposed and underutilized, in a time in which, increasingly, a piece of information that cannot be easily found on the web is assumed not to exist. Even for more persevering hunters, the process takes longer than it could, the risk of missing some piece of vital information is high, and the tools do not adapt well to individual user requirements.

Though some collections, resources and expertise are catalogued and listed in great detail, they can be hard to find, take diverse forms and are often not suited to discovery by potential users.

The aim of this project was to scope new approaches to finding information and collections of relevance to research and teaching at Oxford. It has explored new tools and approaches to enable students and researchers at Oxford and abroad to understand the scope of collections held by the University and to find them quickly and efficiently. It has examined recent developments in the semantic web, linked data and data visualization; considered the application of domain specific tools in other disciplines; and investigated commercial enterprise search solutions to understand the benefits and costs these could bring. In short, this project has sought world-leading solutions for connecting students and researchers at Oxford (and abroad) with the collections that are available to them. Good resource discovery tools, though, are not simply about making research easier and faster, but about facilitating the creation, preservation and discovery of knowledge by enabling new modes of research—especially across disciplines.[2]

Within the overall objective of upgrading the resource discovery facilities of the University of Oxford to the standards which are needed to maintain the institution's worldwide status in research and teaching, this project aimed:

1. to understand how best to enhance the resource discovery capabilities available to members of the University of Oxford, both staff and students, together with other consumers of services offered by the University (such as external researchers admitted to the libraries and museums as well as alumni) so that they are able to carry out their work more effectively;
2. to enhance the global provision of access to resources hosted by the University of Oxford to these groups, enabling their discovery through external tools and thus enhancing the visibility of the University's virtual estate; and
3. to support the University of Oxford in the fulfilment of its Strategic Plan and Digital Strategy by providing tools and services that will support world-leading research and teaching within and across disciplines.

[1] From the University Strategic Plan: Vision http://www.ox.ac.uk/about/organisation/strategic-plan

[2] See The University of Oxford Digital Strategy: http://www.ox.ac.uk/about/organisation/digital-strategy

Scope

The project has taken a wide-ranging view of the meaning of the term ‘resource discovery’. Here it is defined as *any* activity which makes it possible for an individual to locate information which he or she needs. Such material and such activities may be digital or analogue in nature.

In scope

Alongside digital discovery facilities, this means that the scope of this project has included investigation of strategies which do not currently use IT (e.g. printed catalogues and in person discussions) or which may be digital in nature but do not use University of Oxford services (e.g. social media).

Similarly, the scope of relevant stakeholders has been defined widely, including members of the University of Oxford (students, faculty, staff and researchers) external readers, and any members of the public who visit the University museums or access the University’s virtual estate for the purposes of teaching, learning and research. While these users are considered the primary stakeholders here, a secondary group of stakeholders are the myriad departments and individuals tasked with building and managing the current and future resource discovery tools at Oxford. For them, this project aims to provide a unified vision (although not necessarily a single, unified solution) that will facilitate their work within the context of the broader University.

Out of scope

While it is clear that underlying metadata quality is of vital importance to successful resource discovery (which is well-supported in the data collected), the scoping study has not included cataloguing enhancement projects within its recommendations. The recommendations have taken the approach that resource discovery needs to start from the current situation, and not require many years of cataloguing work to be completed before improvements can be made.

Similar remarks apply to fulfilment—that is, the actual access to the information the discoverer wants (e.g., downloading an article after it has been located). The results of this project show that work is needed to improve fulfilment mechanisms — particularly around authentication — but that work is tangential to that proposed here. Again, resource discovery needs to start from the current situation, and not depend on enhancements elsewhere to provide improvements. Essentially, the position to be adopted by the analysis is that resource discovery is the process which comes between data/metadata creation and fulfilment.

Methods and Work: What We Did

This project has conducted 113 interviews, 18 site visits and 3 literature reviews in order to:

- discover the requirements of users of University services (including non-members of the University of Oxford) for resource discovery;
- audit the major local resources which need to be discoverable to these users and their current resource discovery provision;
- investigate the responses of other academic institutions and commercial organizations in the UK and globally with similar requirements; and
- evaluate the current available commercial solutions.

***45** Interviews with users*

***30** Interviews with collection 'providers'*

***22** Consultations with external institutions*

***16** Interviews with vendors/ suppliers*

***7** On-site visits to Oxford libraries and museums*

***3** literature reviews*

While the scoping and analysis has been led by the Bodleian Libraries, it has drawn upon expertise around and outside the University of Oxford on the steering committee, project team, and stakeholder group. The project team has consisted of member of IT Services, the Oxford e-Research Centre, as well as several external consultants with expertise in this area.

The project activities were divided into five areas of activity:

1. **User consultation** (speaking to those in the University and outside to find out what their discovery needs are and how they want them to be provided for)
2. **Oxford collection providers consultation** (speaking to those who professionally guide users to discover the resources they need)
3. **Peer institutions consultation** (recognizing that Oxford is not alone in having resource discovery problems, and seeking to learn from ideas and work elsewhere)
4. **Vendor/publisher consultation** (speaking to providers of software and services which include resource discovery, including search engines, publisher websites and databases, union catalogues and portals, etc.)
5. **Literature search and innovation consultation** (looking at possible sources for innovative ideas which may not be originally intended for resource discovery, as well as the now fairly extensive resource discovery literature)

The data and conclusions from each of these areas is provided in summary form as an appendix, but the overall recommendations have taken into consideration all of the data as a whole.

In total, the project conducted:

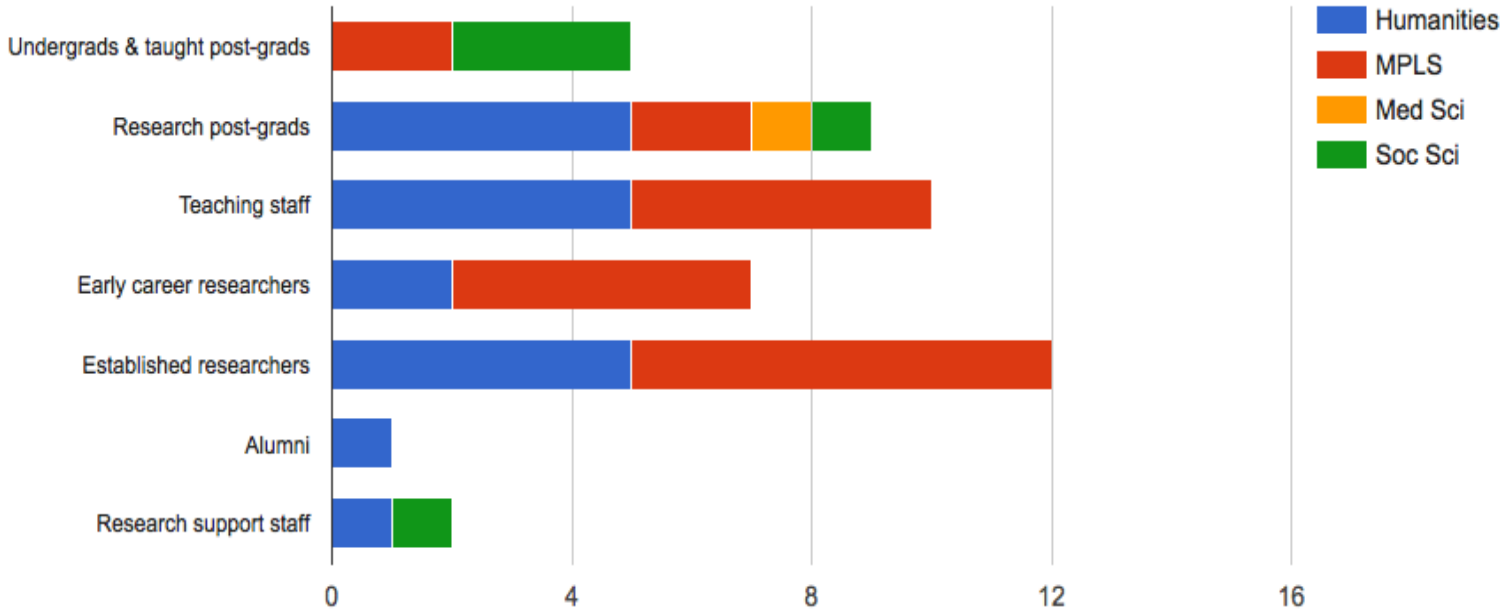
- 45 Interviews with users of collections around Oxford
- 30 Interviews with collection 'providers' (representing all of the collections at Oxford and their users)
- 22 Consultations with external institutions (11 of which were site visits)
- 16 Interviews with vendors/suppliers
- 7 On-site visits to Oxford libraries and museums to observe researchers
- 3 literature reviews

User Interviews

Interviews were conducted with 45 known users of library and museum resources. Faculty were identified using existing personal and professional networks, while students were

identified primarily through a list of volunteers gathered at the 2015 Freshers' Fair. Efforts were made to draft interviewees from all four divisions[3] and try to represent as much diversity in academic / research practice as possible. Medical Science interviewees were the most difficult to recruit, possibly due to greater schedule constraints. The aim for diversity in respondents meant not just looking across the departments but ensuring the selection of people who use a wide variety of research materials. Interviewed users mentioned looking for: printed books and journals, modern papers and archives, manuscripts, museum collections (objects and works on paper), e-books and e-journals, data sets, open access materials, pre-prints, and computer code[4]). The final dispersal of respondents amongst the divisions was:

Figure 1: User Interviews, by division and researcher type



[3] http://www.ox.ac.uk/about/divisions-and-departments

[4] Often from institutional repositories or discipline-based repositories like Arxiv

All interviewers employed semi-structured and person-centred interview techniques. Interviewers began with a structured set of questions, but allowed for significant personalization in responses. Each interview was approximately 60 minutes in length and was recorded in full in order to enhance and substantiate written notes. The interviews were not transcribed in full, but the recordings were re-visited when clarification was required. The data was analyzed by creating a table of responses to each interview question. As patterns emerged from the responses, the table was broken into more and more columns to accommodate more granular coding.

In order to find the 45 interview respondents, the project team approached (most via email) over 92 people. As the interviewees were volunteers, the team recognizes that they were a self-selecting group, who were already likely to be users or supporters of the libraries and museums. When asked for an interview, several people responded that they "didn't use the library or museums at all" and therefore could not be of use to the project. Such respondents were actually sought-after as they provided valuable data about non-use of existing discovery tools and also about perceptions of University resources. The number of interview respondents who did not use existing University finding aids was therefore far below the suspected ratio[5]. In other words, the interview data is skewed towards users of existing library and museum discovery tools and this has been taken into consideration in the analysis of the data by weighting the responses from the 'non-users' more heavily.

The Oxford Providers consultation strand interviewed those from around the University who already have a role in the provision of resource discovery

Oxford Collection Providers Consultation

The Oxford Providers consultation strand interviewed those from around the University who already have a role in the provision of resource discovery, principally those who:

- answer user queries about discovering resources
- provide training to users
- have oversight of work which produces resources with a discovery element

The first way in which this side of discovery was investigated was through exercises to look at the types of queries which experts receive and how they are resolved. The departments involved were the Radcliffe Science Library, Bodleian Special Collections, and two departments from the Ashmolean Museum. The approach was a combination of shadowing and the discussion with the providers of typical queries (using real examples) – the latter because it would not be possible to guarantee that enough interesting queries were received during a few hours on a help desk. The purpose of this was partly to orient the investigator as to the sorts of questions answered, in order to prepare for the later parts of the work.

The second stage, five weeks of interviews, covered archivists, curators, managers and librarians from the Bodleian Libraries, colleges, museums and other academic related departments. In total, 25 interviews were held with 29 interviewees, all of which (except one) were recorded for note taking purposes. These break down as:

- Librarians (13)
- Curatorial staff (5)
- Holders of IT and digital content related posts (9)
- Archivists (2)

[5] To take one example, it is suspected that 20% of students and faculty at the University of Oxford use SOLO.

- Employed by Bodleian Libraries (13)
- Employed by University museums (9)
- Employed by colleges (5)
- Employed elsewhere (2)

Interviews were held in June and July 2015. These were semi-structured in form, allowing concentration on areas of special interest or relevance from the work of each consultee.

Additional information was also gathered from SOLO Live Help session logs and from query counting statistics.

This consultation also included a review and discussions with managers of relevant and related projects. These included:

- ORLiMS - a Staff Innovation Project run by the Social Science Library to facilitate the creation of reading lists;
- Blue Pages - an IT Services and Bodleian project (originally piloted by the Bodleian in 2009) to create a directory of research at Oxford; and
- Oxford Linked Open Data (OXLOD) - a pilot project with the e-Research Centre and the museums to create open linked data.

Peer Institutions

A target list of 23 organizations was selected, chosen in order of preference from the 'Resource Discovery Project Targets for External Consultation' list put together by the Resource Discovery Project Working Group. Marshall Breeding, the author of the 'Future of Library Resource Discovery' white paper, published in February 2015 (http://www.niso.org/apps/group_public/download.php/14487/future_library_resource_discovery.pdf) was also added to the list at a later stage. The selected target list consisted of organizations from the UK, Europe, US and Australia, some of which are similar in size and complexity to Oxford. Of the 24 targets that were contacted, two institutions were unavailable within the project timeframe. The remaining organizations included three museums, two public, two joint libraries and a national archive service (the UK National Archives), two consultants, as well as a wide range of universities.

The main research method was the interview, either in person or via Skype. A list of areas of discussion was drawn up to direct the interviews. This was not envisaged as a rigid list to follow but as discussion points to be tailored to the specific points of interest for each chosen institution to consult. Each interview was recorded (for note-taking purposes only) and a written summary of the interview was sent back to the participants to check and sign off. Whenever possible, efforts were made to speak to several people within the organization, preferably from several divisions (e.g. libraries and museums) to get a fuller perspective. This proved to be difficult given the size and complexity of the participating organizations, the project time-frame and the time of year (the project timescale overlapping with the holiday period).

Vendor/Publisher Consultations

Seventeen providers of discovery tools and services were consulted about their current provision of resource discovery and future plans in the area. This included both commercial and non-commercial providers, of the following types:

- discovery solution providers (both software and cloud oriented) (6 participants);
- collection management software providers (3 participants covering 5 products used or considered for use at the University of Oxford);
- publishers of data for discovery services (who also run their own discovery on their websites) (9 participants);

- managers of union catalogues and services (6 participants)
- bodies providing general digital solutions for teaching and research (1 participant)

Several of the consultees fall into more than one of these categories. A number of other organizations were contacted, and were unable or unwilling to contribute during the project. Consultees included UK, European, Israeli, and US providers.

Consultations took the form of interviews, using multiple communication methods including telephone/Skype calls, email discussions, meetings and discussions as part of seminars.

The discussions were not consistently structured; each being planned around specific objectives relating to the type of provider involved. Many, but not all, of the discussions were recorded for note taking purposes only. Discussions took place between May and November 2015.

Literature Reviews

Three literature reviews were conducted over the course of the project. First, a general review of the literature around ‘resource discovery’ (Appendix 5). This resulted mostly in work produced by or about libraries.

The project also conducted a literature review on innovative areas in information discovery and navigation, with the explicit goals of determining a short list of technologies which might have an impact on resource discovery in the near future as well as identifying potential project partners. This review focused on current research in information retrieval and visualization (Appendix 6). A second literature review was conducted around the use of Social Media for discovery (Appendix 7).

Analysis of Data: What We Found

Users

The interviews with current teaching faculty, research staff, students and alumni uncovered a far more nuanced understanding of search behaviour than is often portrayed in the literature around resource discovery. This project found a number of important patterns which provide the background for recommendations. While a full report on the findings from users can be found in Appendix 1, what is presented here is a brief summary of the findings that had the greatest impact on the recommendations.

Nearly every interviewee said that they rely on asking someone (colleague, supervisor, curator, librarian, or known specialist) at some point in their search process.

This was as true of 'expert' researchers as of 'novice' ones.

Firstly, **resource discovery at Oxford is very discipline-specific**. While quite a few people do start their search at Google, many start at the Bodleian's SOLO catalogue. Within certain disciplines, though, searchers will jump straight to the top resources in their field (arXiv for Physics, PubMed for Medicine, WestLaw or similarly specialized tools for Law). These findings are consistent with the well-documented understanding of the differences in 'known-item' versus subject searching,[6] and emphasize that while that both happen in all disciplines, the sciences are often dominated by the former. One notable exception to this is evidence-based medicine, where researchers are often engaged in very thorough subject-based searching.

There is continued evidence that **students need to learn how to search**. As one professor said, "it's clear that reasonably diligent students are strikingly not sophisticated in their searching. Students search in one place, and if they don't find anything on the first try, they think it doesn't exist."

Nonetheless, **discovery is not as simple as 'novice' vs. 'expert'**. Experts in their fields may use some of the same discovery tools and techniques as incoming students in certain circumstances. A professor in one discipline may, for example, use Wikipedia or basic Google searches to familiarize themselves with a new topic just as a new student might.

Specifically at Oxford, **resource discovery is still a very 'analogue' process for many collections and within many disciplines**. Searchers rely heavily on printed catalogues and hand lists in many different areas as these are the only forms of descriptions about certain catalogues.

Nearly every interviewee said that they rely on asking someone (colleague, supervisor, curator, librarian or known specialist) at some point in their search process. This was as true of 'expert' researchers as of 'novice' ones. **Asking people, and knowing who to ask, seems to make the difference between simply finding what you need to complete an assignment and becoming an expert researcher.** As a senior researcher in Chemistry said "at Oxford, information is power, so nobody shares information online. That's why people are so important to finding information."

---

[6] For a brief description of known-item searching and as short bibliography please see: http://www.iva.dk/bh/Core%20Concepts%20in%20LIS/articles%20a-z/known_item_search.htm Recent discussions about known-item searching have largely been within the context library discovery services and whether they are meeting the needs of known-item searching. See, for example, *Is known-tem searching really an issue for web-scale discovery*? And *Discovery system – testing known item searching* *http://emilysingley.net/discovery-systems-testing-known-item-searching/*

Despite the discussion in the literature around the use of Social Media for resource discovery (see Appendix 7), none of the respondents in this project said that they used open/public social media platforms for asking resource discovery questions. Two said that they have used social media to ask a question, but rarely. Those that do use social media, use it as a way to monitor interest groups, people, conferences, blogs in their field, or as a mechanism to promote their own projects or work. Ten used *Twitter* and seven used *Facebook* for 'keeping up' with developments, including publications, in their field. Ten used *Academia.edu* for this purpose, and four interviewees explicitly mentioned using *ResearchGate*. Students used closed email lists or *What's App/Snapchat* groups to share articles and news items, but this was very much an extension of simply their colleagues in person.

Perhaps most importantly, this project found that **the discovery process was for many searchers about becoming expert in their field as much as it was about finding individual items in their collections.** Multiple interviewees cited 'trial and error' as key to their evolution as researchers. In other words, the more people search and successfully find, the more expert they become in their field. This is in part because they learn the boundaries as well as the tips and tricks for finding the most credible sources and the less well-known parts of the collections. Expertise in a domain requires two things: an understanding of the parameters of your domain and an understanding of the available and relevant resources in those areas. The 'expert' researchers interviewed had varying levels of confidence about their mastery of these domains, but all seem to have a clear sense of its 'borders'. One senior researcher in the humanities used a cartographical metaphor for this: "If I get dropped into the middle of the landscape, I can deduce where I am and navigate my way out, whereas my students will latch on to the first tree that looks interesting."

Outside of Oxford: Vendors and Peer Institutions

Vendors and Publishers

This branch of consultation had a major limitation: commercial organizations concentrate on selling their product(s), and therefore are secretive and biased when taking part in this kind of exercise. Nonetheless, the consulted vendors indicated an increasing interest in:

- Cloud services (many services are now only available in this form)
- Standards and APIs
- Linked open data

Publishers of data were mainly focused on the provision of discovery through their own websites and in the provision of data to aggregate services such as Primo Central and EBSCO Discovery Service (EDS) (though some did not widely disseminate their data to these services); more experimental discovery methods such as Linked Open Data were viewed as outside the scope of their activities. Cleanliness of data, accuracy and availability in the indexes, and use of these services by subscribing customers to provide discovery to the publisher's resources were far more important.

Several vendors, mostly those which already numbered the University as a customer, expressed an interest in collaborative work in resource discovery. Vendors are eager to partner with their customers because they don't know what to do next. Academic and cultural institutions outside of Oxford are largely in the same place: not satisfied with their current discovery tools and looking for 'what is next'.

Peer Institutions

For all the participants in this strand of the research, resource discovery development is very much work in progress, as nobody felt that they had 'cracked it'. Resource discovery at a complex institution requires a lot of resource, especially if there is a lot of customization and/ or local development required. This poses questions of long-term sustainability. Commercial discovery platforms make the job easier but there are limited options for customization, user interface design and further development.

Peer institutions felt that the way forward was with linked data, but that there is no commercial resource discovery system that has fully explored the benefits of this route, so search results tend to be linear and lack in detail. This is no longer sufficient for users who are getting used to a more semantic way of searching on the web. Sometimes not rushing to implement a new system that has come on the market pays off, as things move so fast. There is evidence of rapid improvements in functionality offered by commercial systems.

Looking Forward: The Literature

Three literature reviews were conducted over the course of this project. The first (Appendix 5) provides an overview of literature to date on the general principles of resource discovery. A second focused specifically on the use of social media for resource discovery. A third report looks at the long trajectory of search-based information technologies and discusses opportunities for innovation in resources discovery—particularly in the use of visualization. While the data from users had informed the recommendation in terms of *what to do* this last investigation has significantly informed the recommendations about *how*. Min Chen, Professor of Scientific Visualization at the Oxford e-Research Centre has concluded the report with the following observations:

Although ontology-based search engine technology has been around for about two decades, it has not shown significant effectiveness in resource discovery. The reasons that may explain this problem include the following. (i) It is costly to capture and integrate metadata of resources. (ii) It is costly to support reasonably complex ontologies with necessary techniques such as crawlers and indexing and buffering. (iii) The amount of search activities for library resources is simply insignificant for enabling ontology-learning.

Database-based technology remains the dominant search aid, but its deployment is hindered by the lack of support from visualization for enabling the rapid identification of false positives and false negatives, from interactive visualization for exploring the search space without a set of well-defined search criteria, and from provenance visualization for reducing the cognitive load of remembering what has been searched.

The next generation of technology for supporting the discovery of resources may need to be developed through new innovations while learning the advances of other fields (e.g., online search). It is unlikely that simply borrowing strategy will work. Such an uncreative strategy may actually damage infrastructures and sciences in the long run, as the Internet service providers are using the advantages of the online search technology to take away the services traditionally belonging to library infrastructures. When there is a competition, it is disadvantageous for one party to compete on the other party's term and with the other party's technology.

Interactive visualization and visual analytics should have a significant role in the next generation of resource discovery technology. It is important to understand what visualization is really for. The key is to save the user's time and reduce their cognitive load. However, most of visualizations used in current library technology focus on displaying

search results, while users often find easier and quicker to read the textual results anyway. Hence most existing efforts in this area are unproductive.

Oxford is one of the largest library resource providers in the world. It has the best opportunity to lead the development of library technology through innovation. However, it will always be harder to make and implement a strategic decision to innovate than to borrow.

Recommendations & Next Steps

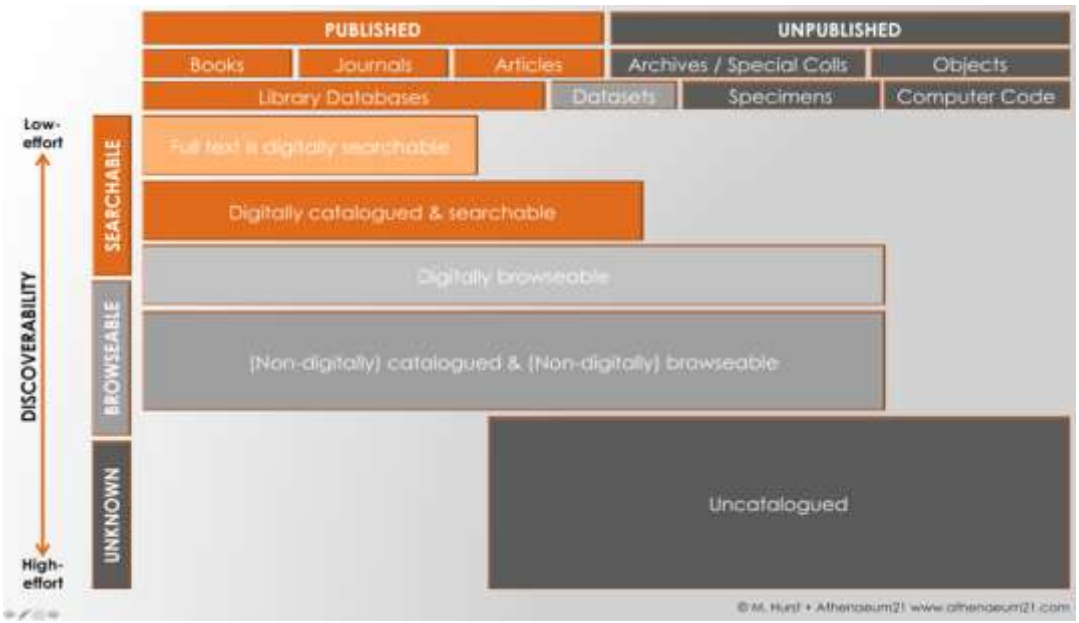
The following recommendations are based on a combination of all of the data and literature consulted over the course of this project. The recommendations are intended to improve the resource discovery situation at Oxford without replacing existing useful tools or attempting to implement a single, monolithic solution.

The recommendations below are based upon the general and well-known understanding of the difference between 'known-item' searching (that is, searching for something specific and known to the searcher) and other forms of searching which are less well-understood. While recommending solutions to improve both types of searching, this project has greatly increased understanding of the latter forms and has uncovered a far more nuanced understanding of how and why people search for resources.

Areas for Investment, Part 1: Mapping the Landscape of Things

The journey from novice information seeker to experienced researchers is not only about gaining skills for searching and finding but about gaining expertise in a particular domain. As mentioned above in the findings, this expertise requires two things: an understanding of the parameters of your domain and an understanding of the available and relevant resources in those areas. 'Experts' have varying levels of confidence about their mastery of these domains, but all seem to have a clear sense of its 'borders'. Therefore, resource discovery should at least in part be about helping people to identify and define these borders.

Figure 2: Types of Materials, and Levels of Discoverability



DPhil students are particularly sensitive to the problem of novelty and domain mapping as they are tasked with finding and creating something original, possibly for the first time in their academic careers. This requires that they first have a firm grasp of what has already been done/ discovered in their field.

Figure 3: Researchers' Discovery of Library Materials: Novice Searchers

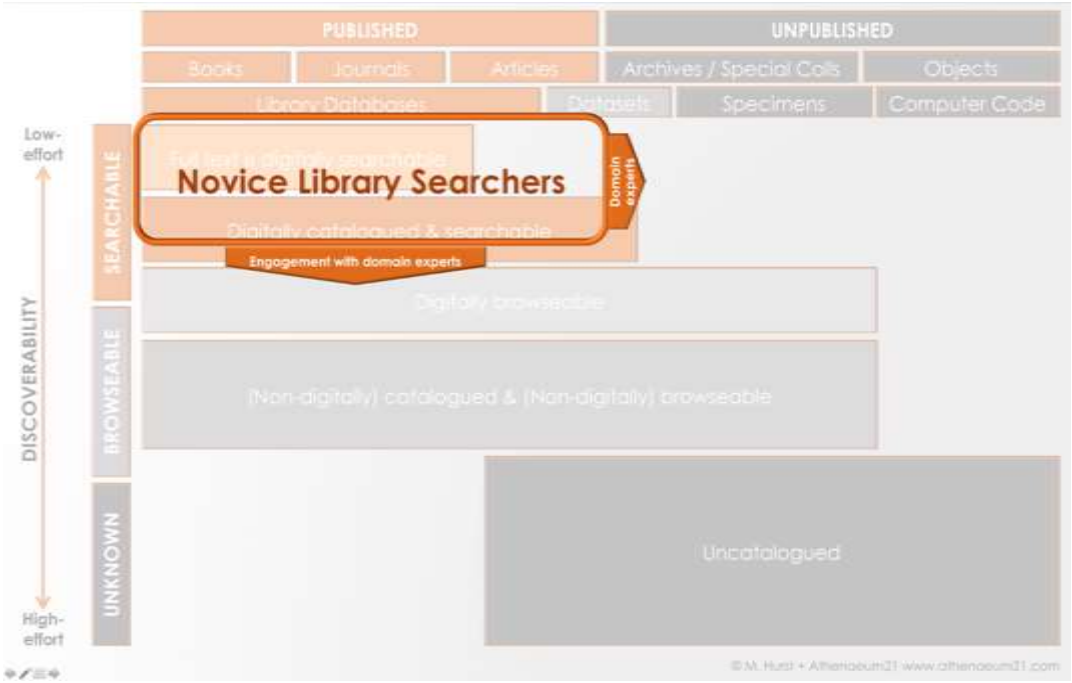
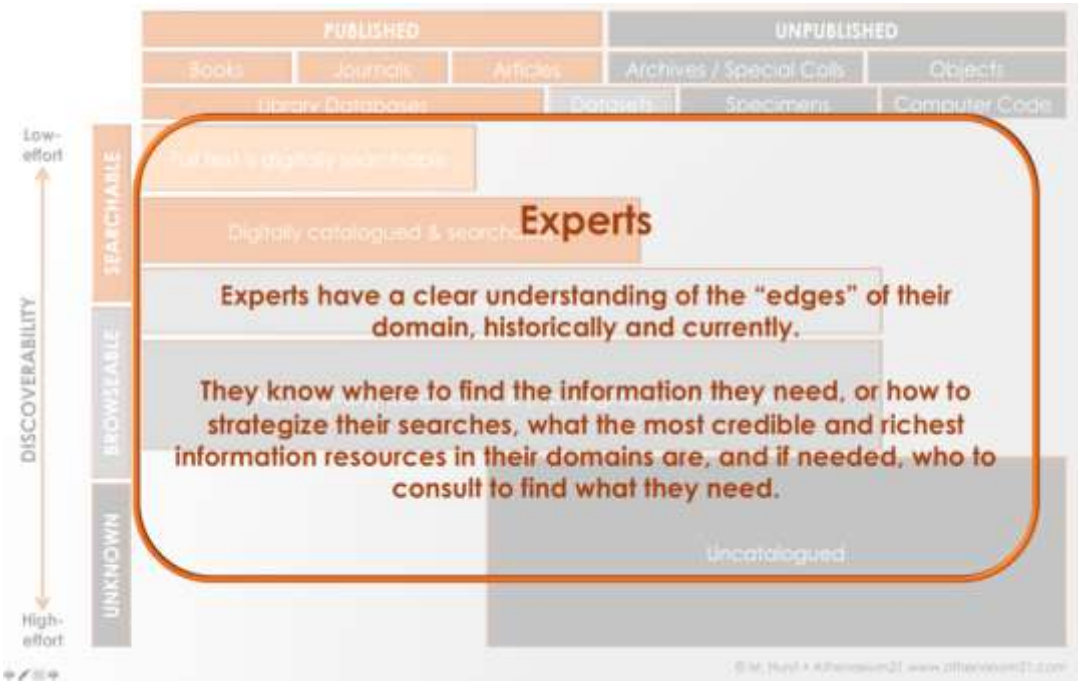


Figure 4: Researchers' Discovery of Library Materials: Expert Searchers

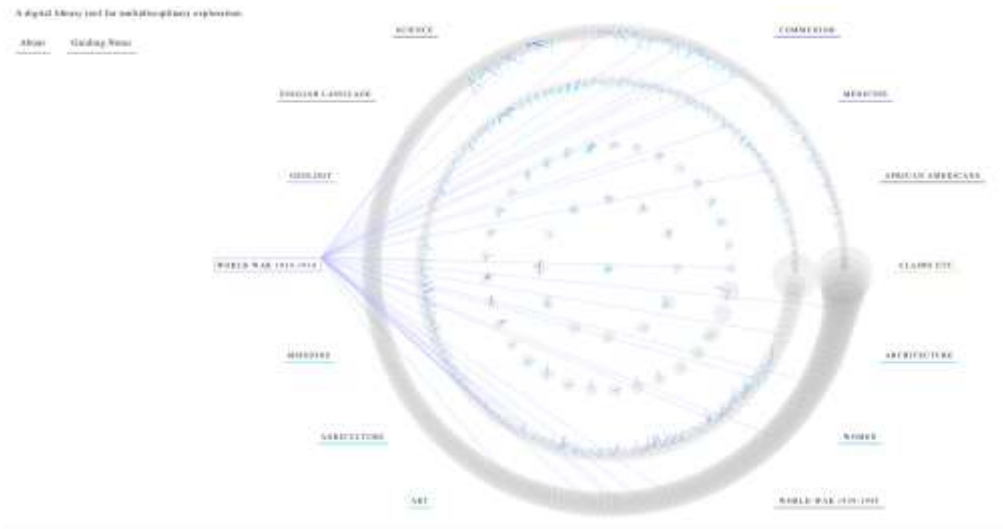


Based on this understanding of information discovery and acquisition, these projects are intended to help searchers of all levels understand both their domains and the information/collections environment at Oxford. These projects are about orienting users to the corpora of collections (digital and non-digital) at Oxford and they include:

Visualizing the scope of the collections at Oxford. Using collection-level metadata, provide an interactive diagram that represents the range of collections at Oxford. Interactive Venn diagrams (for example) could provide a high-level overview of what exists in what form at a basic, or even binary, level. For example, works on paper vs. objects;

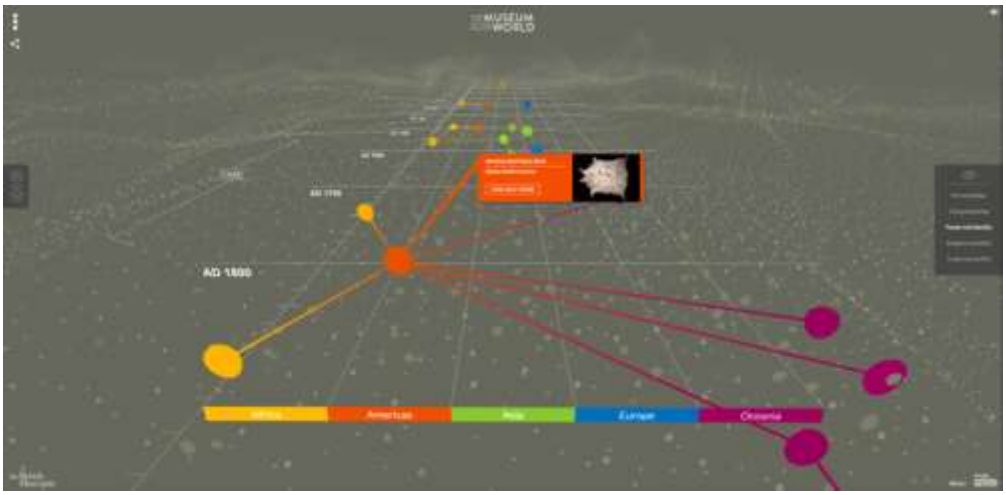
printed vs. manuscripts; visual vs. textual; digitized vs. not digitized; catalogued vs. not catalogued. Dates could be contextualized within ranges of centuries. Such an interface would provide researchers with an immediate visual guide to how many collections are at Oxford and their relative sizes, which ones are searchable electronically, which are catalogued in print indices and which are not yet catalogued. Overlaps in collection provenance, topic or format could provide starting points for navigation. Other views of the data may be able to expose less well-known or unexpected parts of the collections such as that the Pitt Rivers Museum has a manuscript collection and that the Bodleian has art. The *Crossing Disciplines* project at Columbia University[7] (Figure 5) goes some way toward illustrating how such a project might look, but is based solely on bibliographic data.

Figure 5: The Crossing Disciplines project at Columbia University



The British Museum has also explored visual navigation of their collections[8].

Figure 6: History connected: 'The Museum of the World' microsite allows users to explore and make connections between the world's cultures.

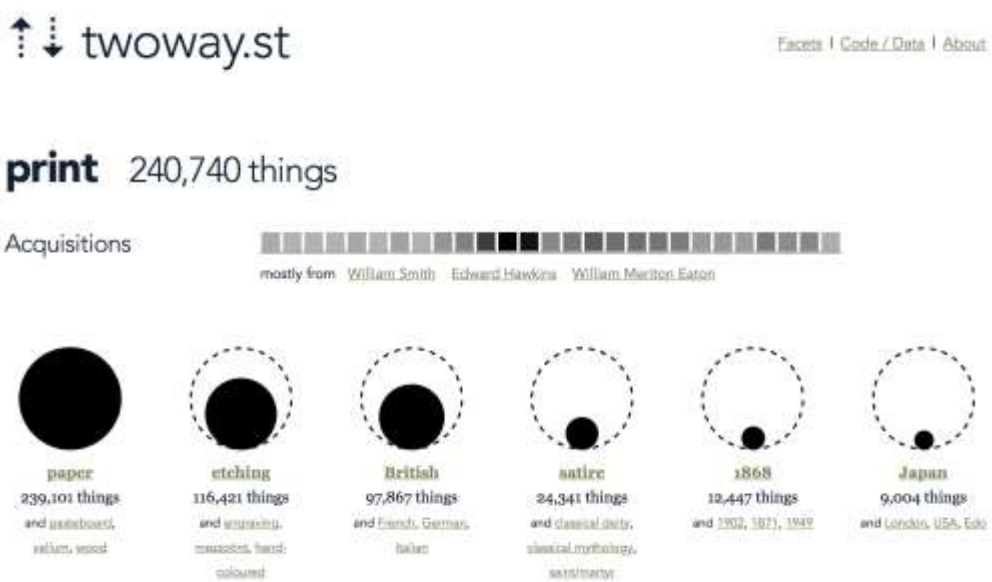


[7] For more information, see http://spatialinformationdesignlab.org/project_sites/library/crossingDisciplines.html
[8] From *The British Museum: A Museums for the World* by Neil MacGregor http://blog.britishmuseum.org/2015/11/12/the-british-museum-a-museum-for-the-world/

'I want to be able to search for Egyptian mummies and get results of articles and books by Egyptologists, the latest scientific analysis of mummies from Chemistry, as well as find where the actual mummies in the museums are'

Cross collection search. Beyond sending researchers to many different places to look for relevant collections, a cross collection search would ideally allow discovery across the libraries, garden and museums, while still allowing users to narrow their search to one or more collections (and allowing the individual collections to brand their own home pages). As one researcher said, 'I want to be able to search for *Egyptian mummies* and get results of articles and books by Egyptologists, the latest scientific analysis of mummies from Chemistry, as well as find where the actual mummies in the museums are'. This project would take existing metadata from across the collections and use a Lucene-based technology such as Elasticsearch to index and expose the existing item-level metadata. This would be intended as an experiment to make the most of existing data without expecting 'perfection' or completeness. Part of the inspiration behind this idea is the project called Two Way Street http://twoway.st/ . Using acquisition data from the British Museum, the project provides an 'independent explorer' for the collections. As described by the project's architect, George Oates, "This is a sketch made quickly to explore what it means to navigate a museum catalogue made of over two million records. It's about skipping around quickly, browsing the metadata as if you were wandering around the museum itself in Bloomsbury, or better yet, fossicking about unattended in the archives."

Figure 7: Two Way Street: a Visual Navigation of the British Museum's provenance data



Taken together, these two projects would provide not only a number of novel ways to navigate the collections, but a clear sense of what is being missed when searching. In other words, the object would be to visualize not only what is being found, but what is *not* being found due to lack of item-level metadata. A professor in the Humanities said, "it's clear that reasonably diligent students are strikingly not sophisticated in their searching. Students search in one place, and if they don't find anything on the first try, they think it doesn't exist." These projects are aimed squarely at addressing that problem.

Expected users.
Based on the data gathered in this project, these tools would be useful for both 'novice' and 'expert' researchers. For incoming students or faculty of all levels, these tools would provide an immediate visual representation of what is available to them and in what format. For seasoned researchers and teachers, this provides the opportunity to explore what they—even after many years at Oxford—may not have known exists.

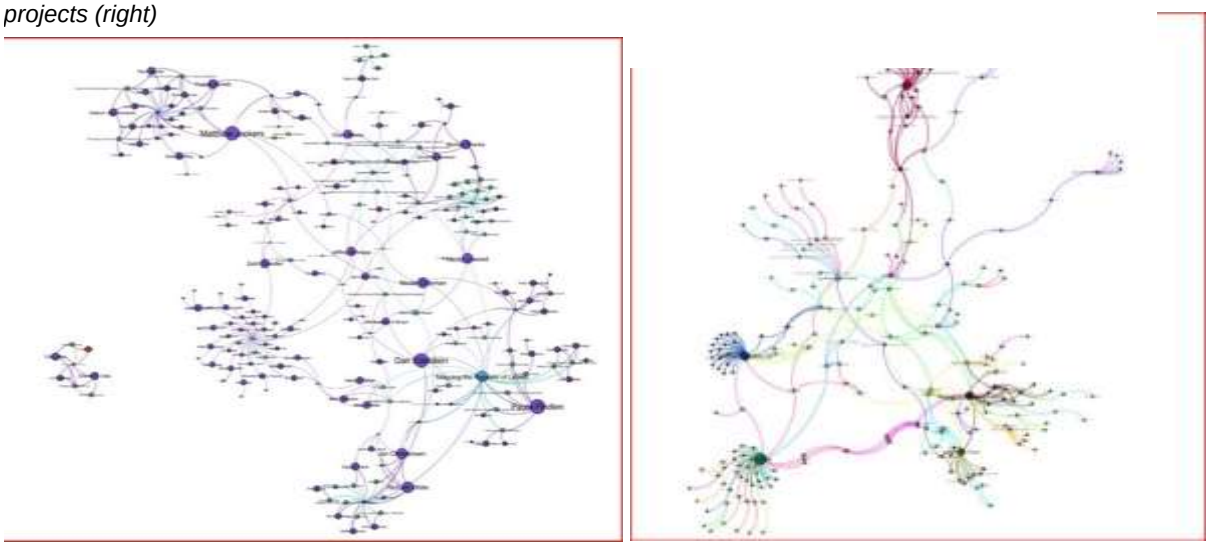
Areas for Investment, Part 2: Mapping the landscape of people.

A second series of recommendations are about connecting *people*. The research conducted indicates that searchers at all levels rely on people—librarians, colleagues, supervisors, mentors or experts in their field—to find resources when other search methods have failed[9] and also when they have not. Asking people, and knowing who to ask, seems to make the difference between finding what you need to complete an assignment and becoming a skilled researcher or even expert in your field. Therefore, these recommendations are intended to help with navigating the landscape of expertise at Oxford.

Create a directory of expertise at Oxford. The current research collaboration tool project (which initially began as the BRII project in the Bodleian in 2009, then became the Blue Pages proejct) aims to provide a directory of research and expertise. Ideally, this project should be built with an open and flexible framework in order to further enable:

Visualization of the network of experts and research. A graph of the professional networks at Oxford would facilitate discovery and navigation within and between fields. It should be noted, however, that a number of institutions have attempted this with varying degrees of success. Care should therefore be taken to learn lessons from others' successes and failures. See the examples below (Figures 8-11) from the Stanford Project on mapping networks of expertise https://dhs.stanford.edu/digital-humanities-at-stanford/dhstanford-gallery/, Harvard Catalyst[10] https://connects.catalyst.harvard.edu/Profiles/display/person/4667/network/coauthors/cluster; and Neurotree http://www.neurotree.org and Academic Tree http://www.academictree.org). There is some evidence from the user interviews that the discipline-specific resources were more heavily-used than the general ones. Academia.edu is heavily used by the interviewees, but mainly as a source of new publications and not necessarily as a means of locating relevant people.

Figure 8: The Stanford Project on Mapping the Digital Humanities can visualize both people (left) and projects (right)



[9] Researchers deem their search to have failed at different points based on their experience and practice. According to their teachers, students will often try one or two keyword searches and then give up. More experienced researchers will have a more sophisticated understanding of when they *should* be finding resources but just aren't.

[10] It would be worth discussing this project with colleagues at Harvard to see if it has been successful. From the 'outside' it does not look very heavily used.

…Particularly amongst students in taught courses, or experienced researchers looking for a particular citation, there are clear things that can be done to make their work easier.

Figure 9: An entry for an individual in the Harvard Catalyst directory

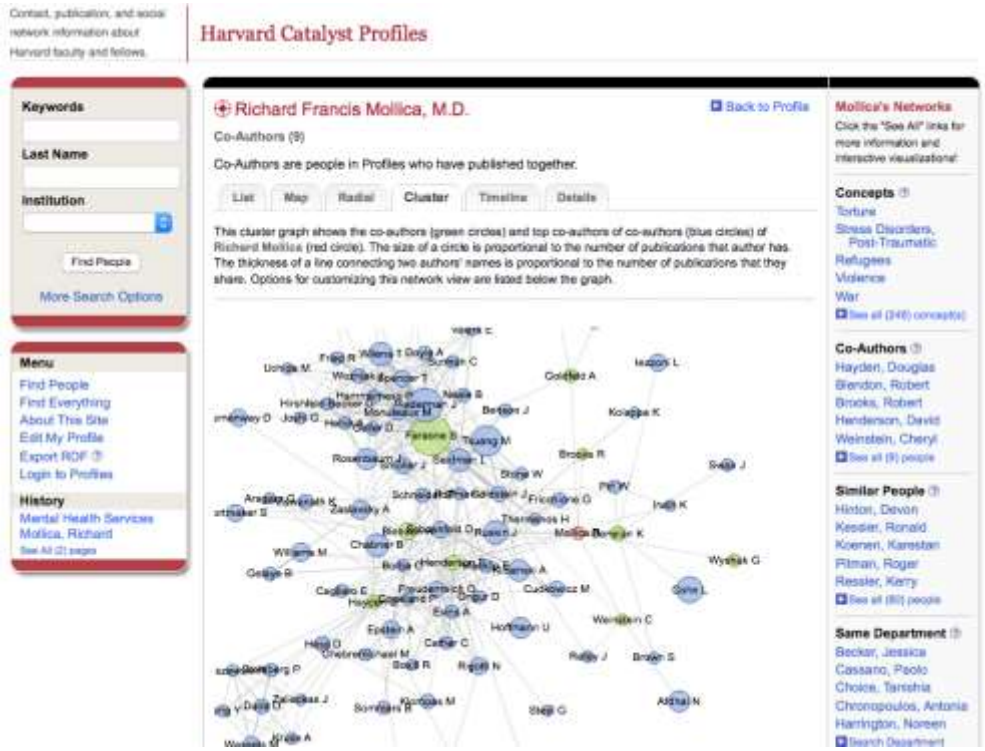


Figure 10: A screenshot from Neurotree, a discipline-specific directory

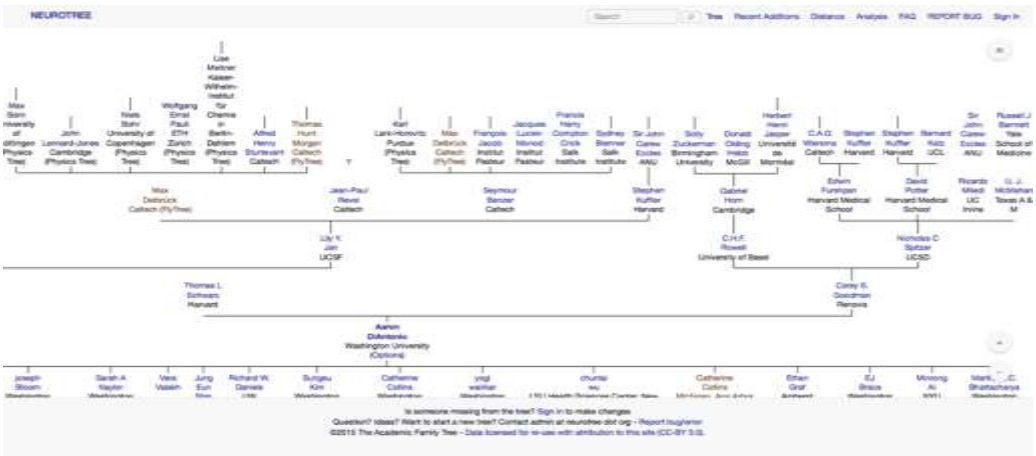
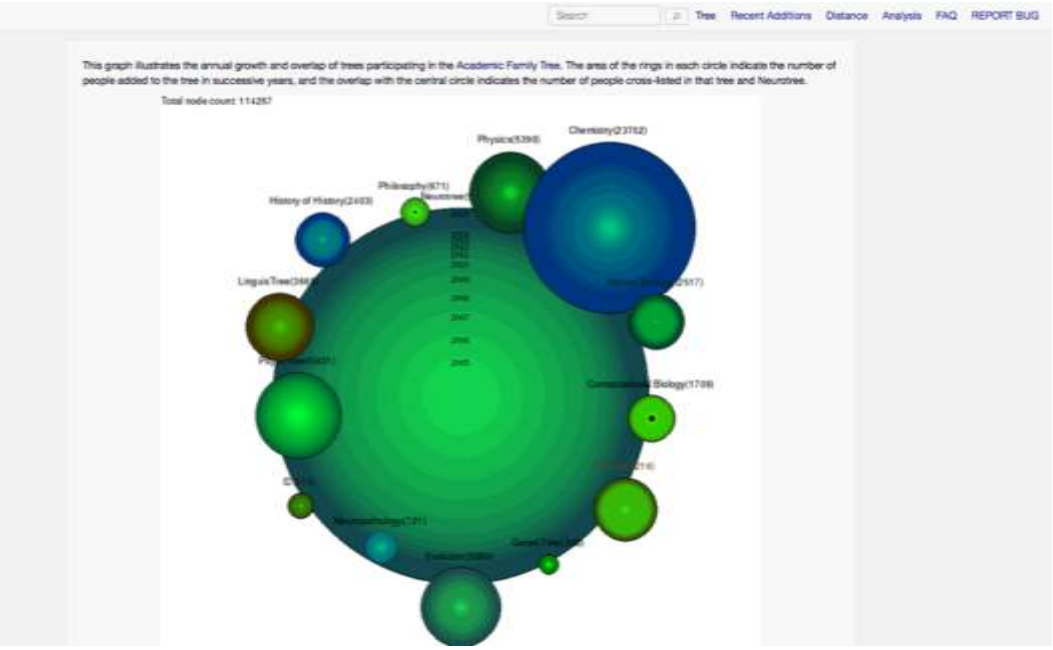


Figure 11: A view of the data in AcademicTree



Connect people, resources, and events.
Several interviewees cited Oxford lectures and talks as being of high value and difficult to find or know about. Providing a reliable source for upcoming talks by division or subject area would be heavily used and well-received. It would contribute to researchers' impact and promote knowledge exchange. Such a system could also be designed to highlight relevant resources to the topics being presented, and highlight Oxford experts (faculty, research groups, etc.)

Expected users.
Asking people (colleagues, librarians, curators) for help in locating resources was universal amongst the users interviewed in this project. These tools were specifically requested by a number of interviewees in the sciences and would be expected to be used for a number of different reasons by a wide variety of users inside and outside of Oxford. As a research chemist said during an interview, "at Oxford, information is power, so nobody shares information online. That's why people are so important to finding information."

Areas for Investment, Part 3: Supporting Researchers' Established Practices

Making research 'faster' or 'more convenient' is a controversial topic. The benefits of serendipity and finding the unexpected are well-documented in the literature and fully supported within this research. For many (across the disciplines), the process of research is as important as the outcome. Nonetheless, particularly amongst students in taught courses, or experienced researchers looking for a particular citation, there are clear things that can be done to make their work easier. These investments will facilitate what is largely called 'known-item' searching.

Getting existing metadata out to the places where many researchers work. Exposing metadata for indexing by Google and Google Scholar would undoubtedly assist those who start their searches on the open web. Working with subject-specific repositories like arXiv and PubMed, and publishers like JSTOR would further assist in connecting users with specific Oxford resources, particularly those who say "I always start at Google."[11]

Finding an available electronic resource and not being able to access it was a frustration to a number of users, though, so exposure of metadata on the open web should go hand in hand with **improvements to authentication systems.** A number of users expressed frustrations at finding things through Google and not being able to access them and at the variation in search results when working from different locations. As a DPhil student in Medicine said, "I wish someone had told me when I started my degree why I should use the VPN connection. I get totally different search results." Most interviewees realized that they got different search results in Google and some other databases from different locations, but didn't understand why. An improved authentication experience and education around the differences between authentication options would support users—particularly in their use of Google.

Facilitate Citation Chaining. "I find 1 or 2 articles and then I just go from there." (Research post-graduate in the Social Sciences). Citation chaining is ubiquitous in all areas of research across all disciplines. CrossRef and other individual tools and databases have gone some of the way towards making this easier, but is still not heavily present. Searchers in all disciplines use cited references as authoritative points of departure for finding more resources on a topic.[12]

Expected users.
In contrast (and complementary to) the tools for mapping landscapes, these projects would facilitate precise 'known-item' searching by anyone looking for a specific work or citations. Making metadata available to those who start their searches on the open web also facilitates engagement by those outside of Oxford.

Laying the Foundations

Undergirding all of these recommendations are a set of general principles and recommendations. These are to be thought of both as means of moving forward with the recommendations of this report and as a set of principles for best practice.

Moving Forward

Shared discovery will necessitate shared infrastructure of systems, people and policy. This should be supported by **establishing an interdisciplinary research group** to take forward the recommendations of the report. Participants should include the museums, libraries, Oxford Internet Institute, the e-Research Centre, IT Services and other universities. The group would need an executive to ensure momentum and delivery.

---

[11] The University should also be aware of the tradeoffs of exposing its metadata to search engines in this fashion. The metadata will be used and re-used by commercial companies and that may be a fine trade-off for enhanced findability, but this should be entered into knowingly.

[12] Much as with 'saving time' citation chaining is quite controversial. It can create an echo-chamber effect where the same few sources are cited over and over. This appears to be the case both with published materials and with archival collections.

Recommendations & Next Steps

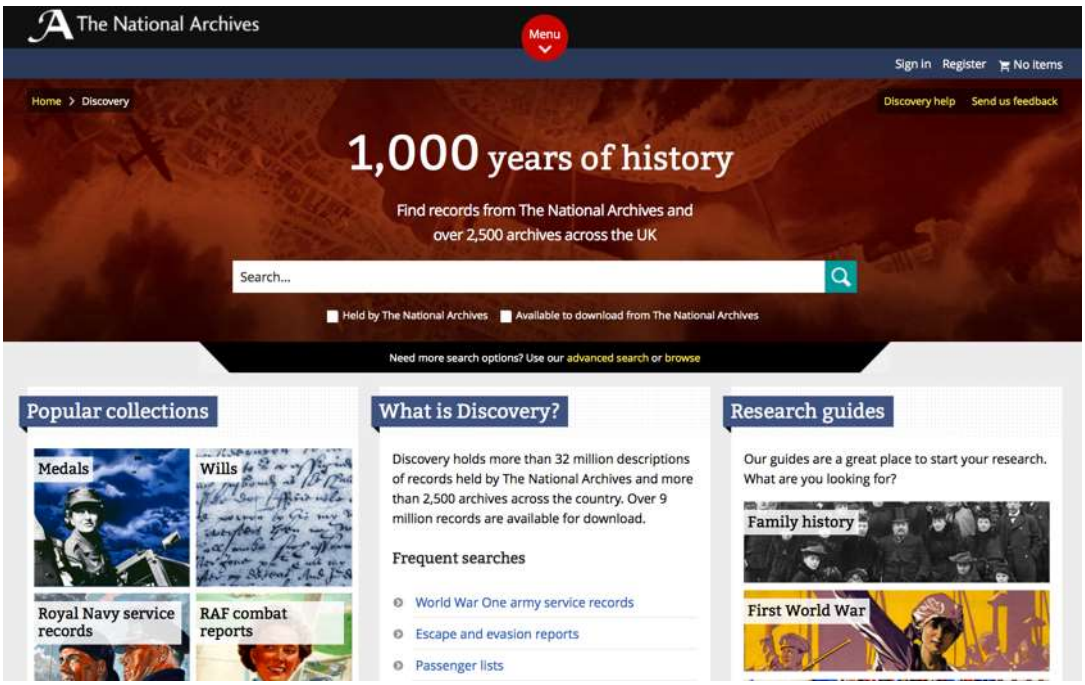
Embedding this work in research will support **investment in the analytics and data infrastructure to support evidence based decision making across the collections**. The project made it clear that collection managers actually have very little data about how their tools and collections are being used and yet this data is essential for assessing the impact of current and future innovations.

The Academic Services and University Collections (ASUC) should investigate the creation of **a 'Collections @ Oxford' portal** that would provide access to:

- Cross-collection search
- 'Navigate the Collections'
- A platform for digitized content throughout the University
- New Scholarship @ Oxford (linking to ORA and OUP)

The UK National Archives Discovery page[13] provides a good example of portal for discovery services across diverse collections. It combines a cross-collection search with prominent featured 'popular collections' as well as research guides.

Figure 12: Discovery at the National Archives



A number of **existing and proposed projects** will help move the University toward the recommendations in this report, including:

- Researcher collaboration tool project
- Oxford Linked Open Data (OXLOD)
- Platform for digitized content throughout the University

[13] For more information see: http://discovery.nationalarchives.gov.uk

Best Practice

- **Collaboration** within the University and with partner institutions around the world is essential to providing innovative solutions for discovery. Collaborators should be sought with similar academic institutions but also in private industry and 'bleeding edge of information design (see Jonathan Harris, Greg Hochmuth, George Oates, and others).

- **Digital is not always faster, so it needs to have some added value**. "I used the digitized card catalog, which is just pictures of the cards. It was so imprecise. You had to flip through it. It took forever. I would have rather physically handled the cards, and it would have been faster. Digital is not always faster." (Archives researcher in the Bodleian.) Physical card catalogues are faster than browsing images of cards online. There are many reasons for digitizing collections but projects should be based on a firm understanding of desired outcomes and the added benefit brought by the digital.

- **Complete collection metadata**. To state the obvious, collections cannot be discovered using electronic search tools unless they have some sort of representative electronic description. High quality description is key to discovery, and funding should be made available to improve metadata and to better understand strategies for prioritization of description.

- **Thorough and relevant training and education around discovery.** Users who participated in museum or library inductions or specialized training services, were unanimous in their assessment of their value in helping them both find what they need and improve their information skills.[14] Negative comments were largely about conflicts in scheduling and availability. One recommendation to increase attendance at such sessions could be to re-brand the training from being about 'how to find things in the collections' to 'how to become an expert in your field.' Partnering with the Open Access and ORA teams to train researchers (for example in how to write an abstract that will make your research findable) in the entire scholarship/research lifecycle may further enhance participation.

[14] The researchers here admit there is a likely bias in the data collected as the interview respondents (by virtue of their willingness to be interviewed) were likely to be library and museum enthusiasts and/or supporters.

Benefits

"So we create these little capsules of collections—examples of things—and if we aren't giving the students a way to branch out from there, a way to see more of something, then we aren't doing anyone any favors."

(Teaching Fellow in the Museums)

A shared strategy for discovery will:

- **Benefit teaching, research, and public engagement by facilitating knowledge exchange around collections**. This is in direct support of the University's Priority 1, "Global Reach" and helps to realize the importance of public engagement (Point 10 of the University Strategic Plan)
- **Support interdisciplinary research by showing the overlaps in collections and expertise at Oxford** (In support of the University's Priority 2: Networking, communication, and interdisciplinarity in recognition of point 13, "Many of today's research questions cut across traditional subject boundaries." This will further support the University's Digital Strategy aim 1, to "Enable new modes of research especially across disciplines")
- **Mean that students, staff, and external researchers will not have to understand how the University is organized in order to find what they are looking for** (Commitment 6. "To ensure that the unique richness of the collegiate University's academic environment is both retained and refreshed." Providing an IT Infrastructure that allows for a more seamless experience across University resources will enable a better student experience as outlined in the University Strategic Plan, Enabling Strategy 4, points 72 and 75).
- **Make the most of what we have by making use of existing metadata.** (Making efficient use of the resources at hand and partnering to fulfil some of the identified needs, should free up the resources for other areas. (Enabling Strategy 1)).
- **Stimulate new research.** In support of the Commitment 1, "To maintain originality, significance and rigour in research within a framework of the highest standards of infrastructure, training, and integrity")
- **Aid in discovery of collections that are not yet digitally cataloged** (University Strategic Plan, Point 40. "Strengthening Oxford's global and digital online presence, as signalled in our new priorities, will ensure students studying at Oxford have improved access to materials.")
- **Enable innovation across the collections by exposing metadata for use by machines as well as people** (Supporting point 9 in the University Strategy" An enhanced online presence" and "Promote new ways of generating, curating, and engaging with data (e.g. visualization, data analytics, digital collections, augmented datasets, health)" from the Digital Strategy.
- **Allow the Libraries and Museums (and Colleges?) to provide mutual technical and policy support for one another** (Enabling Strategy 4 "To invest in information technologies that enhance the capacity of Oxford's academic communities to collaborate with each other and with global partners, and that support the student experience.")
- **Enable new modes of research and collaborations by making prominent some of the overlaps in both collections and expertise** (Point 16. "Sharing resources, advanced facilities, and collections is a useful way of developing research interdisciplinarity through bringing individuals together who have common interests." and Point 15. "Oxford has long pioneered multidisciplinary degrees. New thematic research collaborations will lead to new study opportunities for undergraduates and postgraduates.")
- **Combat the problem of 'if it isn't digital, it doesn't exist' by visualizing the gaps in digital metadata.** (Point 20. "The maintenance of a sustaining research environment is crucial to the University's research standing. We will enhance the infrastructure which supports research at the highest level, including libraries, laboratories, museums, and information systems.")

Postscript:
A Resource Discovery Dystopia

by Professor David De Roure

As long as the centuries continue to unfold, the number of books will grow continually, and one can predict that a time will come when it will be almost as difficult to learn anything from books as from the direct study of the whole universe. It will be almost as convenient to search for some bit of truth concealed in nature as it will be to find it hidden away in an immense multitude of bound volumes. When that time comes, a project, until then neglected because the need for it was not felt, will have to be undertaken. — Denis Diderot, "Encyclopédie" (1755)

'It will be almost as convenient to search for some bit of truth concealed in nature as it will be to find it hidden away in an immense multitude of bound volumes.'

Back in 2000s everyone was talking about "information overload", fearing more data, more channels, more interruption. Discussion declined with the rise of personalisation—in search results, in social media recommendations. Filtering was the answer, as in Shirky's 2009 "It's not information overload. It's filter failure". By 2011 the risks of personalisation had been acknowledged, with the realisation that academics were living in "filter bubbles" where algorithms gave them what they wanted—and sealed them off from new ideas. This led to community "echo chambers" where ideas were amplified and self-reinforced by the penumbra of relevant information.

By 2015 the situation had deteriorated further. The digital marketing ecosystem had blurred with scholarly resource discovery. Algorithms made recommendations not just on reading habits but to maximize publishing corporations' income—purportedly to fulfil their legal obligation to shareholders' interests. In distorting information discovery, it corrupted the very process of research, bringing accusations of algorithmic censorship of scientific work. Universities were not far behind the publishers in privileging the academic outputs that would maximize their research income through optimising locations and citations, an increasingly automated process. A competitive ecosystem of algorithms arose—culminating in the infamous 2020 "Russell Algorithm" court ruling mandating transparency in search technologies.

Research was increasingly conducted *in silico*, and rising automation increasingly bypassed the human—especially as rewards went to those first past the post with research results. As machines became the predominant producers and consumers of research content, they became the primary users of real-time resource discovery algorithms. While liberating for humans, this also denied critical reflection and challenge—they only got to see what machine-learning algorithms, and those who configured their behaviour, wanted them to see. Research methods had become built into our knowledge infrastructure, unchallenged.

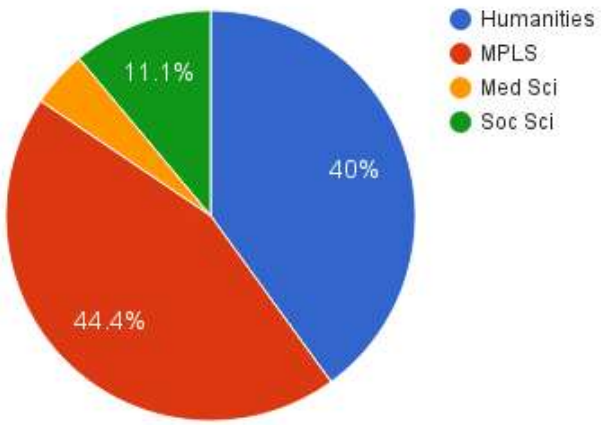
A small band of academics and librarians started meeting in secret…

Appendix 1:
Summary of Data from User Interviews

by Christine Madsen & Megan Hurst, Athenaeum21 Consulting

Interviews were conducted with 45 known users of library and museum resources. Faculty were identified using existing personal and professional networks, while students were identified primarily through a list of volunteers gathered at the 2015 Freshers' Fair. Efforts were made to draft interviewees from all four divisions[15] and try to represent as much diversity in academic / research practice as possible. Medical Science interviewees were the most difficult to recruit, possibly due to greater schedule constraints. The aimed diversity meant not just looking across the departments but ensuring the selection of people who use a wide variety of research materials. Interviewed users mentioned looking for: printed books and journals, modern papers & archives, manuscripts, museum collections that included text (including coins, tablets, and plaques), visual works (both objects and works on paper), e-books and e-journals, data sets, open access materials, pre-prints and computer code.[16] The final dispersal of respondents amongst the divisions was:

Figure 13: Distribution of User Consultations, by Division



Efforts were also made to select participants at all levels of their research activities and careers. This resulted in participants across seven different researcher 'types': undergraduates and taught post-graduates; research post-graduates, teaching staff, including lecturers and tutors; early career researchers; established researchers; alumni; and research support staff. Interviewees often fell into more than one category within this

[15] http://www.ox.ac.uk/about/divisions-and-departments

[16] Often from institutional repositories or discipline-based repositories like Arxiv

Appendix 1:
Summary of Data from User Interviews

matrix—someone could, for example, be both an alumnus and a senior researcher. Therefore, in the analysis of the data, respondents were assigned both a primary and a secondary 'type' category.

Figure 14: User Interviews, by division and researcher type

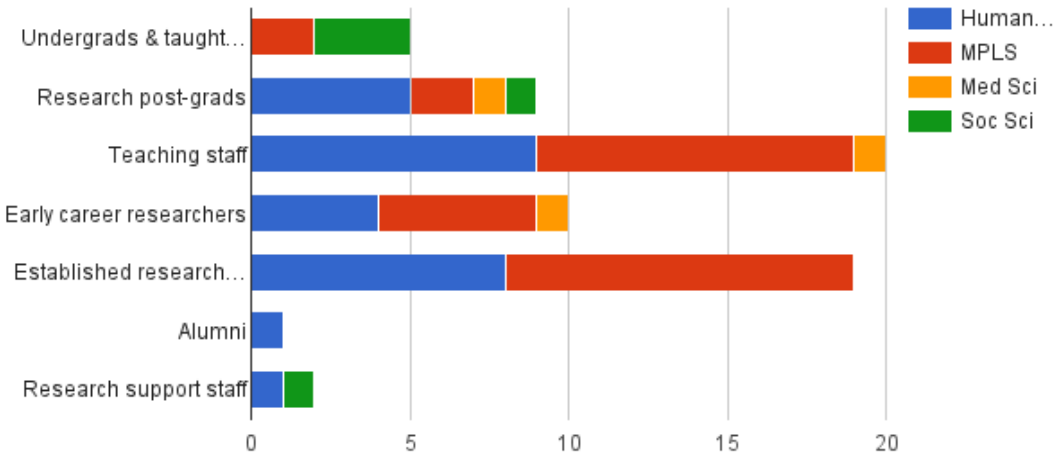
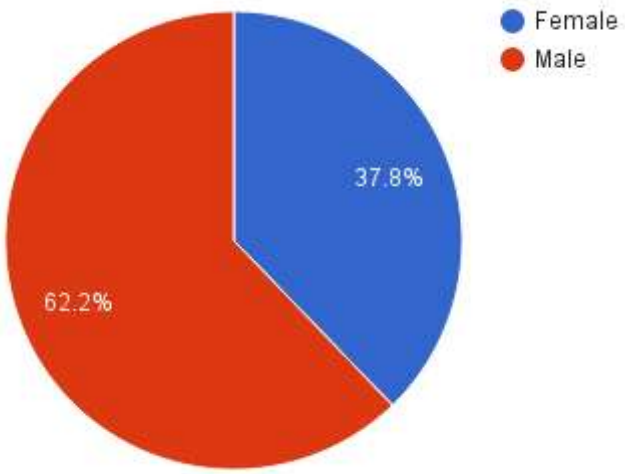


Figure 15: User Interviews, by gender



Appendix 1:
Summary of Data from User Interviews

Figure 16: User Interviews, by age range

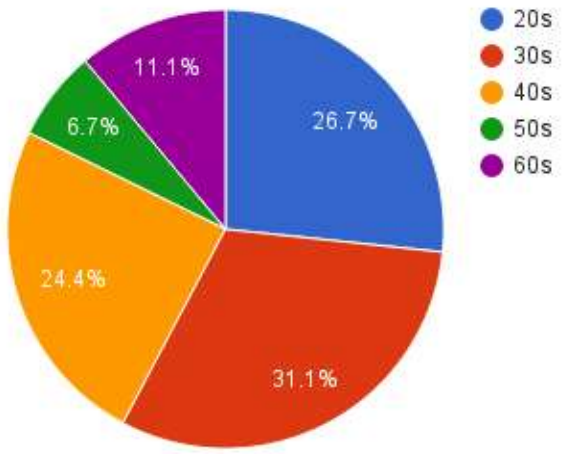
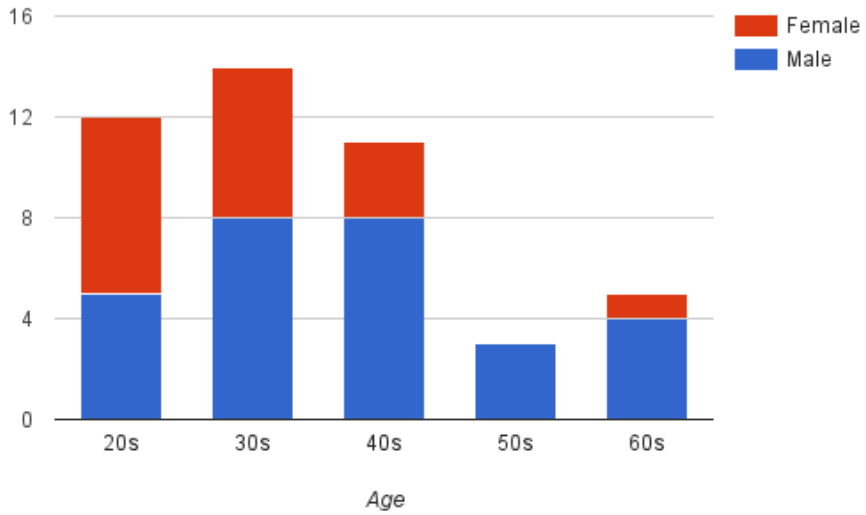


Figure 17: User Interviews, by age range and gender



All interviewers employed semi-structured and person-centred interview techniques. Interviewers began with a structured set of questions (see below), but allowed for significant personalization in responses. Each interview was approximately 60 minutes in length and was recorded in full in order to enhance and substantiate written notes. The interviews were not transcribed in full, but the recordings were re-visited when clarification was required. The data was analyzed by creating a table of responses to each interview question. As patterns emerged from the responses, the table was broken into more and more columns to accommodate more granular coding.

In order to find the 45 interview respondents, the project team approached (mostly via email) over 92 people. As the interviewees were volunteers, the team recognizes that they were a self-selecting group, who were already likely to be users or supporters of the libraries and museums. When asked for an interview, several people responded that they “didn’t use the library or museums at all” and therefore could not be of use to the project.

Such respondents were actually sought-after as they provided valuable data about non-use of existing discovery tools and also about perceptions of University resources. The number of interview respondents who did not use existing University finding aids was therefore far below the known ratio. In other words, the interview data is skewed towards users of existing library and museum discovery tools and this has been taken into consideration in the analysis of the data.

Analysis of Data from Users

Users

The interviews with current teaching faculty, research staff, students and alumni uncovered a far more nuanced understanding of search behaviour than is often portrayed in the literature around resource discovery. This project found a number of important patterns which provide the background for recommendations.

Disciplinarity

Firstly, resource discovery is very discipline-specific. While quite a few people do start their search at Google, many start at the Bodleian's SOLO catalogue. Within certain disciplines, though, searchers will jump straight to the top resources in their field (ArXiv for Physics, PubMed for Medicine, WestLaw or similarly specialized tools for Law). The responses between the interview respondents were nearly equally divided between those who started their searches with SOLO, Google or Google Scholar, or a subject-specific resource. These findings are consistent with the well-documented understanding of the differences in 'known-item' versus subject searching and emphasize that while that both happen in all disciplines, the sciences are often known to be dominated by the former.

Specifically at Oxford, resource discovery is still a very 'analogue' process for many collections and within many disciplines. Searchers rely heavily on printed catalogues and hand lists in many different areas as these are the only forms of descriptions about certain catalogues.

Information Skills and Literacy

This research also provided continued evidence that students need to learn how to search. As early as 2008, the UCL CIBER report on the 'Google Generation'[17] pointed out the striking discrepancy between facility with information technology and the ability to find and assess information online amongst the so-called digital natives. This research indicates that not much has changed in the last seven years and good information-seeking skills and information literacy (the ability to discern good sources from bad) need to be taught. As one interviewed professor said, "it's clear that reasonably diligent students are strikingly not sophisticated in their searching. Students search in one place, and if they don't find anything on the first try, they think it doesn't exist."

Nonetheless, discovery is not as simple as 'novice' vs 'expert'. Experts in their fields may use some of the same discovery tools and techniques as incoming students in certain circumstances. A professor in one discipline may, for example, use Wikipedia or basic Google searches to familiarize themselves with a new topic just as a new student might.

[17] This project produced a number of outputs, which are summarised in a presentation here: http://www.webarchive.org.uk/wayback/archive/20140614113419/http://www.jisc.ac.uk/media/documents/programmes/reppres/gg_final_keynote_11012008.pdf

Appendix 1:
Summary of Data from User Interviews

The Role of Training

Most teaching faculty did not directly teach searching skills (although some did). They relied on students to have gone to library inductions to learn basic search skills such as Boolean logic. Those who did teach their students would provide very discipline-specific skills just as foreign-language searching or how to work with specific disciplinary tool (e.g. Early English Books Online).

Most of the 14 students interviewed had been on library induction training for their course. None of the interviewed students had been on a museum induction. These were usually subject-specific inductions provided by a subject-librarian (law or history, for example.) Feedback on training was positive, (although we recognize that our data was skewed in favour of 'library supporters') with every interviewed respondent saying that the training was helpful.

More specifically, these inductions seemed to have a very direct impact on where people were starting their searches and how they were searching. For example, four students mentioned using OxLip+ for their searches, in direct response to the library training they received. Three people mentioned not using 'Articles and More' for the same reason. The interviewees that participated in library induction training did also appear to have gained the ability to distinguish credible from non-credible information (e.g. the limitations of Wikipedia, and to look for .edu and .gov sources when conducting open web searches).

Feedback on SOLO was generally positive, although the data indicates that it was largely used for 'known-item' searching. For students, this was usually in response to citations on a reading list or a name and title mentioned in a lecture. This was substantiated by the on-site observations at four libraries (Radcliffe Camera, RSL, SSL, and the Law Faculty), where the SOLO search histories were recovered on public terminals. Of the 30 searches recovered, all but one were a combination of title words and part of an author's name. In the majority of these cases, the (apparently) correct search results were retrieved. Cases of nil or incorrect results were due to mis-spelling (usually of the author's name) and the user re-typed the name and received the correct result.

Figure 18: A typical example of a search with no results

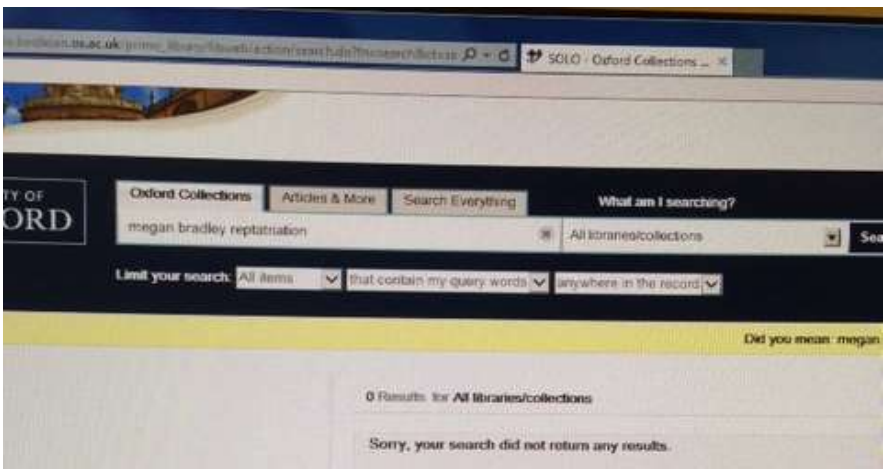
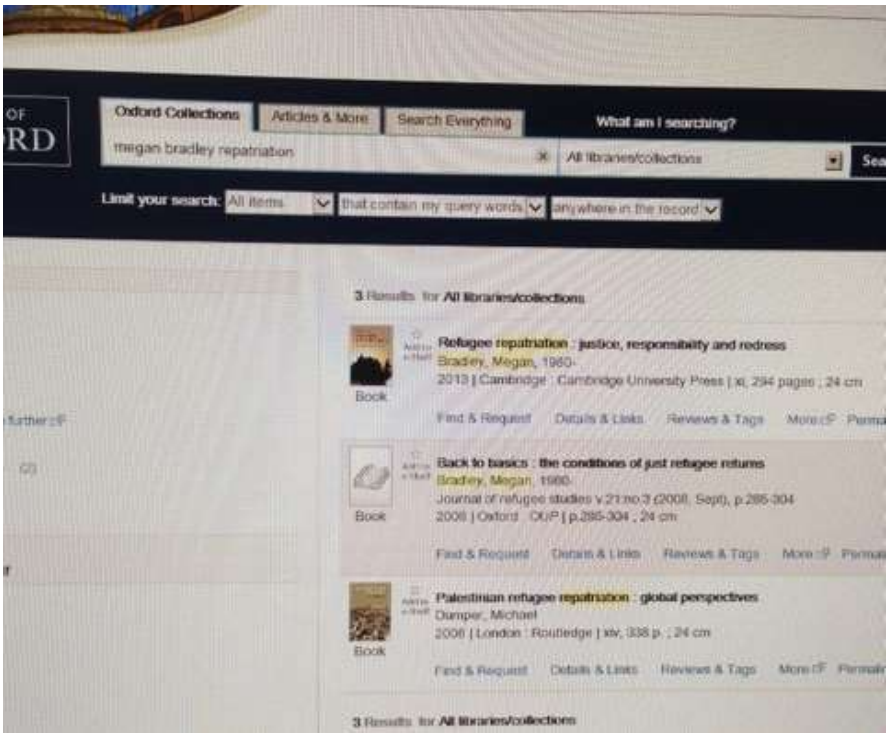


Figure 19: The same search with the spelling mistake fixed



The Role of People

Nearly every interviewee said that they rely on asking someone (colleague, supervisor, curator, librarian or known specialist) at some point in their search process. This was as true of 'expert' researchers as of 'novice' ones. Asking people, and knowing who to ask, seems to make the difference between simply finding what you need to complete an assignment and becoming an expert researcher. As a senior researcher in Chemistry said "at Oxford, information is power, so nobody shares information online. That's why people are so important to finding information."

Social Media and 'Keeping Up'

The few respondents that had a sense that they were able to 'keep up' with the new publications and research in their field attributed this to the narrowness of their field and small size of their global research community. The vast majority of respondents (even those known experts in their field) did not have high-confidence that they were "on top of" everything that was happening in their domains. When asked specifically how they keep up in their field many directly responded, "I don't." Of the incoming students (both graduate and undergraduate) very few tried to monitor new publications and were mostly responding to suggestions from supervisors and instructors. For more senior academics, most had developed mechanisms for coping with 'keeping up'. These usually involved a combination of social media, informal communications (email from colleagues), conferences, Zetoc and table of contents alerts from specific journals, and/or more formal roles such as serving as editor or reviewer for relevant journals.

Despite the discussion in the literature around the use of Social Media for resource discovery (see Appendix 7), none of the respondents in this project said that they used open/public social media platforms for asking resource discovery questions. Two said that they have used social media to ask a question, but rarely. Those that do use social media, use it as a way to monitor interest groups, people, conferences, blogs in their field, or as a mechanism to promote their own projects or work. Ten used *Twitter* and seven used

Facebook for 'keeping up' with developments, including publications, in their field. Ten used *Academia.edu* for this purpose, and four interviewees explicitly mentioned using *ResearchGate*. Students used closed email lists or *What's App/Snapchat* groups to share articles and news items, but this was very much an extension of simply their colleagues in person.

The Role of Expertise

Perhaps most importantly, this project found that the discovery process, for many searchers, paralleled their evolution as experts in their fields as much as it was about finding individual items in their collections. In other words, the more people expand the boundaries of what and where they are searching (with tacit and explicit assistance from mentors in their domains, and often through their own trial and error), the more expert they become in their field because they learn the boundaries as well as the tips and tricks for finding the most credible sources and the less well-known parts of the collections.

Expertise in a domain requires two things: an understanding of the parameters of your domain and an understanding of the available and relevant resources in those areas. The 'expert' researchers interviewed had varying levels of confidence about their mastery of these domains, but all seem to have a clear sense of its 'borders'. One senior researcher in the humanities used a cartographical metaphor for this: "If I get dropped into the middle of the landscape, I can deduce where I am and navigate my way out, whereas my students will latch on to the first tree that looks interesting."

Devices

Of the twenty-two interviewees that were asked which devices they use to conduct research, the vast majority use laptops as their primary device, with three using desktop computers, and one using an iPad. Of the laptop and desktop users, seven used iPads and tablets as secondary devices for reading at home, or while in transit. Five people said they occasionally used their phones for Google queries or 'quick searches' while in a lecture, or as a 'last resort.' One researcher said that she uses her phone to search SOLO while walking around the library.

User Interview Questions

1. Can you briefly describe your area of research/course of study/teaching area?
2. Can you describe a typical research task?
    a. What kinds of materials are you generally looking to find?
    b. How do you know where to start?
3. What are your 3-5 most heavily used resources for finding materials.
    a. (if they answer with all electronic discovery tools, then which is their favourite and least favourite and why?)
4. How do you keep up with the work in your field?
    a. How do you find out about new publications?
5. On what type of device do you usually do this work?
6. Where do you do your research? From home, office, library?
7. Did you ever have any training —either here or at another institution —in how to find resources?
    a. How was it? Useful?
    b. How did you know / learn how to become a researcher?
8. Can you show us a typical 'research task' that you have done recently? Talk us through the decisions you are making
9. Do you see anything on the horizon that may change how you do your work / find things?

For teaching faculty

Appendix 1:
Summary of Data from User Interviews

10. How do you instruct your students to find things in their area of study?
    a. do you send them to specific resources?
    b. do you give them ‘instructions’ on how to search?
11. Do you provide any training to your students?
    a. or do you recommend certain training for your students?
12. How do you share resources with your students? (eg, printed lists, online reading lists, websites, etc.)

Appendix 2:
Summary of Data from Oxford Providers

by Ray Stacey

Interviews were held with 30 people inside the University of Oxford (see Participants below), whose current roles are somehow tied to the management or provision of collections at Oxford. These participants were people who have a role in the provision of resource discovery, principally those who:

- answer user queries about discovering resources
- provide training to users
- have oversight of work which produces resources with a discovery element

In this way, these interviewees can be said to have some sort of expertise in the existing methods of resource discovery at the University.

These interviews covered all forms of collections: museums, libraries, datasets, digitized collections, and born digital collections. The interviews were structured around a list of designated topics and questions (see 'Areas for Discussion' below). These are some of the key points from the series of consultations carried out with Oxford Experts.

Library Specific

- Librarians prefer to use the individual search in each application rather than an overall search for example the search in a specific database rather than from within SOLO.
- Ideal is to use Google and library resources; it is not sensible to use a single search to find everything.
- Libguides – useful to place material for teaching sessions which also help to promote libguides. Difficult to edit and add material, it is very clunky. Lots of text and not designed. Not easy to add videos, twitter or images.

Museum Specific

- The museums are overwhelmed by the volume of enquiry emails received. It might be useful for them to investigate ticketing (enquiry management) systems so that the enquires can be more easily managed.
- Aggregated systems, at the national level, have with open or two exceptions died. If aggregated systems are one of the solutions then it would need to be compulsory, have a common thesaurus and semantics and require no additional resource requirements for the museums and libraries.

Resource Discovery

- Lack of electronic cataloguing and digitization will hold back Resource Discovery. All the museums and libraries I visited commented that they don't have enough resources to do more digitation. If they did not receive more funding they would struggle to put resources into Resource Discovery.
- How to find a collection - usually need to find an expert. One of the most important resources that users need to find are people.
- When setting up websites for collections need to consider what the use cases are. There also needs to be more consistency across the websites and this could be achieved by defining a standard framework across the museums and libraries.
- Resource Discovery must be intuitive for users - i.e. shouldn't need too much help to use.

- Need to ensure that any resource discovery doesn't only show what we have digitized but links to catalogues for those objects not digitized and to an expert for any resource not even catalogued. There isn't a digitization issue at the Bodleian and the museums but a resource description issue.
- Education is key, Resource Discovery must include education.
- Don't allow technology to dominate what we do.
- The more I put collections online the more queries I receive.
- Mobile access to very important and must be able to navigate the collections on a mobile data. There has been a massive increase in access from mobile devices. Will the use of devices such as smart watches make this worse and also need to consider the attention span especially of children.
- Google is the elephant in room for Resource Discovery. It is used by almost everyone, even though it often isn't suitable, and if a resource isn't found in Google they think it doesn't exist. Also any search operating in a database or across a collections needs to work in the same way as Google, otherwise users may struggle to use it.
- Cataloguing collections and archives is very different from cataloguing books and journals. Could the first step for Resource Discovery be 'just' for collections and archives with books and journals added at a later date?

Areas of discussion with internal interviewees

Information to aim to gather before meeting/discussion

1. Generic institutional information (e.g. size, departments, affiliated organizations etc.)
2. Library / Museums / IT structure (e.g. convergent service, library service structure)?
3. Institutional information environment (e.g. library collections, museum & special collections, archives, research repositories, VLE etc.)
4. Current resource discovery mechanisms & products (including some or all of library, museums, archives and special collections, finding people/experts)

Discussion points (to be tailored to the specific points of interest for choosing the institution to consult)

1. Background
    a. What is your role at the library/museum/college?
    b. Who are you customers/clients? Who do you provide services to?
    c. What sort of queries do you have? and how are they resolved?
2. Current resource discovery structure
    a. What discovery tools do you use and/or recommend?
    b. Why do you use these particular discovery tools?
    c. How would you improve the discovery tools you use?
    d. What don't you like about these tools?
    e. Do these tools meet your needs and the needs of the customers/clients?
    f. User education - how are users encouraged to choose the right tool, discovery strategy etc. (especially how to cope with differing discovery needs)
    g. What would help you most? more choices, better signposts, how would you improve user education?
    h. How could any solution delivered by this project lighten your load?
3. Where do your queries come from? What are the most popular routes?
    a. the institutional website,
    b. the library website/ portal,
    c. museum website,
    d. direct access to a resource or discovery via subject databases or publisher sites,
    e. Google scholar/ Microsoft Academic Search/ other non-library routes

Appendix 2:
Summary of Data from Oxford Providers

4. Future plans
    a. Do you have any procedural or other changes planned?
    b. If you had an unlimited budget what would you do?
5. Other
    a. Anything relevant not covered by the above.
    b. Social media

List of Consulted 'Providers'

Cathy Scutt plus others from Social Sciences Library	Education Subject librarian, Bodleian Library
Charlotte Goodall	Assistant Librarian Classics, Bodleian Library
Michael Riordan	Archivist Queens College and St Johns College
Anna Sander	Archivist Balliol College
Hannah Chandler	SOLO Live Help Organizer and Official Papers Librarian, Bodleian Library
Christine Marsden	Head of Digital Programmes, Bodleian Library
Sally Rumsey	Digital Research Librarian, Bodleian Library
Jonathan McAslan	Electronic Resources Manager, Bodleian Library
Alison Felstead	Head of Resource Description, Bodleian Library
Martin Poulter	Wikimedian in Residence, Bodleian Library
Paul Trafford	Web Officer, History of Science Museum
Jeremy Coote	Curator and Joint Head of Collections. Pitt Rivers
Jill Walker	Assistant to Director and IT officer , Botanic Gardens
Adam Marshall	ORLiMS project, IT Services
Kirtsy Taylor	Head of Library and Information Services, Green Templeton College
Karine Baker	Psychology, Physiology and Anatomy Subject Librarian , Life Sciences/ Medicine librarian, Bodleian Library
Catherine Hartley	Reader Services Librarian - St. Anne's College
Jerome Mairat	Curator of Heberden Coin Room, Ashmolean
Paul Collins	Assistant Keeper for the Ancient Near East, Ashmolean
Lotte Boon	Head, Research Systems and Information Management Team, Research Services
Mark Dickerson	Librarian. Pitt Rivers
Darren Mann	Head of Life Collections, Oxford Museum for Natural history
Francesca Leoni	Yousef Jameel Curator of Islamic Art, Ashmolean
Anjanesh Babu	ICT Assistant (Networks), Ashmolean
Elizabeth McCarthy	Communications, Bodleian Libraries

Appendix 3:
Summary of Data from Peer Institutions

by Masha Garibyan

Purpose

The Resource Discovery Project at the University of Oxford aims to deliver long-term improvements to resource discovery at the University, both outward and inward. The first stage of the project involved analysis and planning of the work to be done. One of the strands of the project was to consult external organisations who were either considered similar to the University of Oxford or who had been doing interesting work in resource discovery.

Research methods

The main research method used was interview, either in person or via Skype. Each participant was given a consent form to sign prior to the interview. All contact details and progress notes were saved in an Excel file for future reference. Each interview was recorded (for note taking purposes only) and a written summery of the interview was sent back to the participants to check and sign off. Whenever possible, efforts were made to speak to several people within the organisation, preferably from several divisions (e.g. libraries and museums) to get a fuller perspective. This proved to be difficult given the size and complexity of the participating organisations, the project time frame and time of year (the project timescale overlapping with the holiday period). Supporting evidence (e.g. user surveys, notes etc) was sought for each interview but was not always available.

Participants

A target list of 20 organisations was selected, chosen in order of preference from the 'Resource Discovery Project Targets for External Consultation' list put together by the Resource Discovery Project Working Group. Marshall Breeding, the author of the 'Future of Library Resource Discovery' white paper, published in February 2015 (http://www.niso.org/apps/group_public/download.php/14487/future_library_resource_discovery.pdf) was also added to the list at a later stage. The selected target list consisted of organisations from the UK, Europe, US and Australia, some of which are similar in size and complexity to Oxford. Of the 21 targets that were contacted, two institutions were unavailable within the project timeframe. The remaining organisations included three museums, two public libraries, two joint libraries and a national archive service, as well as a wide range of universities.

A list of areas of discussion was drawn up to direct the interviews (appended to this report). This was not envisaged as a rigid list to follow but as discussion points to be tailored to the specific points of interest for each chosen institution to consult.

Research findings

Resource discovery tools used by participants

The participating organisations use a number of single search resource discovery platforms, including:

- Primo
- Summon

- VuFind
- EDS
- OCLC WorldCat
- Blacklight
- Home-grown systems

One institution has decided against having a resource discovery platform and is concentrating on supporting users with their discovery needs, regardless of where discovery starts. They are in the process of phasing out their OPAC. .

Two institutions use several discovery tools to support different functions of their single search platform (e.g. Summon, EDS & Primo). This practice appears to be relatively common in the US. Blacklight has been used by several institutions to link together several resource discovery tools to provide a single interface.

Most institutions have kept their OPACs, although some institutions are planning to phase them out in the near future to direct most of discovery through the single search platform.

The majority of institutions have an Open Access repository, typically DSpace, ePrints or Hydra. Some institutions are in the process of acquiring or implementing a research data management system, mostly a commercial solution.

Material coverage

The extent to which material is included in the single search varies, depending on the size of local holdings, human resource availability, political decisions etc (e.g. whether the associated museums wish to have their finding aids indexed in the resource discovery platform). For large & complex organisations, integrating numerous finding aids is a real challenge and requires considerable time and effort. Not all the institutions have currently integrated their archives and special collections, either for political reasons or because it is still work in progress. Not all the institutions have enabled an article-level search in their single search discovery system, some have only made a journal title search available.

Some organisations use unified discovery platforms (e.g. Summon) to provide access to a particular type of resource (e.g. journal articles), rather than multiple resource formats.

Reasons for choosing current discovery services/ products

The participants identified the following main reasons for choosing a particular discovery platform:

- A comparative study of the options available at the time
- Commitment to open source solutions
- An existing client of a vendor that developed a single search discovery platform
- No existing commercial system that fulfilled the organisational needs

Comparative analysis of different resource discovery systems

As mentioned above, several participating institutions conducted a comparative study of web scale discovery systems available at the time.

Two institutions mentioned that the comparative study did not reveal a clear winner, that all the main systems on the market did a reasonable job, so it was easier to go with the vendor that already had an existing relationship with the institution. Several institutions also

mentioned that they didn't think it was worth the effort to change over to a new system unless it offered something significantly better than the existing one. Paul Stainthorp, in his article titled 'How commercial next generation library discovery tools nearly got it right' (http://paulstainthorp.com/2011/05/17/how-commercial-next-generation-library-discovery-tools-have-nearly-got-it-right/), published in 2011, argues that 'the differences between [web-scales services] are not that significant....thinking that ...there are some 'good' and some 'bad'...is probably wrong. It's not really about the product, it's about the willingness of the vendor to overcome problems, and about their attitude to their customers'.

The results of a confidential lightweight trial of several web-scale discovery systems (EBSCO Discovery Service (EDS), Summon, Metalib+ and Google Scholar) kindly shared with the consultation by one institution indicate overall that the four search and discovery systems are very similar. However there are differences in terms of relevancy and other issues. In the categories of relevance ranking and successful full-text linking, the differences among EDS, Summon, and Metalib+ were statistically insignificant and no one system greatly outperformed the others. Google Scholar fell behind in both categories. In the category of up-to-date results, the disparities were slightly greater though not significant enough to warrant a recommendation, with EDS scoring best, followed by Metalib+, then Summon, with Google Scholar falling greatly behind the others in this category, with the exception of searches in the Sciences, where it scored higher than the other systems for up-to-date results. Known-item searches are the only area where Google Scholar significantly outperformed the other systems across disciplines.

A small-scale Primo usability study (5 experienced researchers) was also shared with the consultation:

- Not as interested in Web 2.0 features, eg Tags, as had been thought previously
- Advanced search box not liked
- Didn't understand the word 'facets', preferred the word 'filters'
- Would start their search at such places as Google, Wikipedia (to only to get general background) and sites such as Web of Science, rather than the library page

There are a number of external online resources that can offer a useful overview of the most popular library discovery systems on offer, for example:

- Christison, A. (2013). Discovery layers and discovery services. *Catalogue and Index*, 170, 2-12. http://shura.shu.ac.uk/7435/1/AC170_article.pdf (open access)
- Unified Resource Discovery comparison chart (Google, open access) https://sites.google.com/site/urd2comparison/home/comparison

Important user requirements involved in decision/ customisation

Most participants have done some user consultation prior to acquiring or developing a single search resource discovery system. A variety of consultation methods have been used, e.g. lab testing, interviews, online feedback, focus groups, etc. Below are some of the user considerations identified by the participants:

- User needs partly depend on a particular discipline (e.g. a Humanities scholar vs. Scientist)
- Optimisation for Google and other search engines, as both local and external studies have shown that a high percentage of discovery (including licensed e-material) starts outside of the library, e.g. via Google or Google Scholar.

- Need to concentrate on the majority of user needs, can't do it all. The 'longer a library tries to do everything 100% right, the higher the risk that it will lose users'.
- Technical considerations have to come after full analysis of user needs and not the other way around.
- Discovery tool promotion is according to the relative value identified for different disciplines.
- Some organisations have created broad user categories to help identify user requirements

Comments on individual resource discovery platforms

Summon

- Good for searching journal articles
- Good presentation of results, users can evaluate the initial results without going into the individual resources
- Out-of-the-box installation was clunky, search wasn't reliable, so had to work closely with Summon to identify and implement enhancements. Further changes need to be made to make it easier for users to access material.
- Limited flexibility

VuFind

- A strong support community
- Not much customisation is needed to get it to work
- Very flexible
- Continued improvement & development is possible ('grows with you')
- Tags can be used to create lists of items of interest to a user or related to a particular course or reading list

Primo

- Can handle a lot of different data formats
- Good relevancy ranking
- Does 'find' and 'identify' pretty well
- Gives a good selection of search facets
- Users need to do too many clicks to get to the resource
- Found it hard to configure Primo to expose records in Google
- Would be nice to be able to do a subject-specific interface
- Not good at dealing with local holdings
- Limited customisation options
- Linear structure, no 'navigate'/ semantic search function
- Search results are not hierarchical - not good for providing context, linking records together. This is a particular issue with archives & special collections

- Collection facet doesn't work very well for Archives & Manuscripts (although some of the issue is gaps in cataloguing)

Blacklight

- Services multiple record formats
- Advocacy of open source technologies
- Greater control over the discovery interface
- Content is crawled by Google by default (MARC records)
- Active development by a US-based group of institutions to add new features
- Good community support system
- High level of flexibility
- Limited fulfilment capability, no ability to move across systems to get access to content, unlike Summon, where this feature works quite well, – not easy to evaluate the initial results without going into the resources

Home-grown museum system

- A standard-based approach to developing museum systems, so that they can be shared with the museum community
- A more interactive way to engage visitors, taking into account typical user behaviour

Home-grown archive system

- A unified presence
- Copes with wide variety of data (eg 'born digital' records, websites, data)
- Flexible and friendly interface design
- Re-using data, lots of potential
- Object oriented architecture helps develop the system further, e.g. developing APIs
- Having a single search box is familiar to users from their Internet/ Google experience, so it enables users who aren't familiar with how archives work (and who may not be comfortable using archives otherwise) to benefit from the service
- Responsive design, works well on portable devices

Satisfaction with the current resource discovery set-up

There was no single organisation that felt that their resource discovery set up was perfect, each organisation could identity a number of areas that required improvement or/ and further development, e.g.:

- Reducing the number of steps for authentication for external access to resources
- Presentation of sets of resources (e.g. archival and print material) on particular topics
- Simplification of content and navigation to web pages
- Better use statistics
- Better record harvesting by Google

- Further work on bringing various solo catalogues (the library catalogue, repository, archives) together to make it easier to index data

One institution commented that while they thought that it would be great to have a single search interface for all the discovery tools, it's a challenge to do it in such a way that it's still clear and doesn't look cluttered.

Another institution would like to create a new generation discovery architecture that would 'have a layer capable of searching the metadata, constructing really smart things like GIS, date interfaces, name interfaces of the companies, displaying photographs on a rotating basis etc' but have struggled to find an individual with enough creative vision to drive the project forward.

Linked data/ semantic search

Traditionally, providing context for material was seen as more important for archival holdings but this has changed. Simply providing access to a single resource is no longer sufficient. Librarians and archivists are working closer together in the community to tackle this issue. The linked data, CIDOC CRM (http://www.cidoc-crm.org) and other developments are also playing part here.

The ResearchSpace project is developing a collaborative environment for humanities and cultural heritage research using knowledge representation and Semantic Web technologies. The **CIDOC Conceptual Reference Model (CRM)** provides definitions and a formal structure for describing the implicit and explicit concepts and relationships used in cultural heritage documentation. One of the problems for researchers, in terms of research data, is that there is usually no context and, therefore, no stable basis for building knowledge. The CIDOC Conceptual Reference Model (CRM) transforms raw data and embeds and makes explicit, as part of the process, the implicit meaning that is not stored in traditional database formats. In Humanities and Cultural Heritage, context is particularly important but difficult to integrate due to the differences in perspectives, objectives, and the history of the particular organisation that produced the data. The Semantic framework concentrates on allowing these perspectives to exist harmoniously rather than using traditional methods of unified catalogues where different datasets are squeezed into a unified model.

http://www.dlib.org/dlib/july14/oldman/07oldman.html - Oldman et al (2014) Realizing Lessons of the Last 20 Years: A Manifesto for Data Provisioning & Aggregation Services for the Digital Humanities (A Position Paper), D-Lib Magazine, July/August 2014, Volume 20, Number 7/8 [Online]

One of the challenges of the semantic model is that technical work has to happen alongside curatorial work, as context creation requires specialist knowledge. The archival community is developing a conceptual model (probably more akin to FRBR than CIDOC CRM) in addition to the existing cataloguing standards, but for the archival sector. This work is being led by the International Council on Archives (http://www.ica.org/13799/the-experts-group-on-archival-description/about-the-egad.html).

Library website issues/ terminology

Several organisations mentioned a recent library web re-design project to improve library presentation of discovery tools, navigation etc. All the institutional participants could identify areas of improvement for their library websites.
One institution took an interesting approach of agile website development where they select a small area of improvement and make a series of changes in response to user feedback before arriving at the final version.

Appendix 3:
Summary of Data from Peer Institutions

There are issues related to terminology used to describe the tabs used in a single search interface.

Here are some examples:

- 'Everything', 'Online', 'Physical'
- ‘SuperSearch’, ‘Catalogue Search’, ‘Full text E-journal Search’
- ‘Everything’, (OPAC), ‘Articles’
- ‘Books & more’, ‘Articles & databases’, ‘Databases A-Z’, ‘Journals’
- ‘Search Articles +’

Libraries often use ‘more’ or ‘+’ (as in ‘Books & more’) to describe tabs, which can be confusing to users (more of what?).

All the participants use simple search as default. Users are now used to the Google model where ‘people do a search first and then filter afterwards’.

Users tend to use the default tab the most, regardless of what it is.

There are two ways in which search results are presented: a single list and a Bento box. Blacklight and VuFind customers use the Bento box approach, as it is part of the platform design. According to one university, the Bento box is popular with institutions that preferred a home-grown system. Another university mentioned that external research showed a 50/50 split for user preferences for a single list vs. the Bento box approach. When they were investigating the options, there was no clear winner. The choice of approach seems to largely depend on the underlying solution, e.g. Primo sites have a single list and Blacklight sites go with the Bento box.

User education

LibGuides is the most popular tool for producing subject and other library guides.

Several institutions mentioned that the use of their single search interface largely depends on how it is being promoted by academic staff, local museums and branch libraries.

User education provided by museums and public libraries is quite different from user education provided by universities. It is quite low key in comparison. Typically users are given information when they visit the library/ museum and/or look at the organisational website.

One museum consulted took the unusual approach of ensuring that backend staff have a frontline duty rota, so if a visitor needs help with using the museum mobile app or have some other technical query, they can get help from the Museum technical experts. This approach is very popular with visitors and gets a consistent high performance ranking from visitors.

Approaches taken to address different discovery routes used by patrons

There is general agreement amongst the participants that a significant proportion of resource discovery starts outside the library, mainly in Google or Google Scholar. This is the primary reason why Utrecht University decided to concentrate its efforts on providing discovery support for users irrespective of where discovery starts, rather than investing significant resources and effort in acquiring and maintaining a resource discovery platform.

Appendix 3:
Summary of Data from Peer Institutions

Utrecht was inspired by the 2009 **Ithaka Faculty Survey** (http://www.sr.ithaka.org/sites/default/files/reports/Faculty_Study_2009.pdf).

For example:

“Since the first Faculty Survey in 2000, we have seen faculty members steadily shifting towards reliance on network-level electronic resources, and a corresponding decline in interest in using locally provided tools for discovery.” (p. 4). Also note the declining use of library buildings and catalogues as starting points for research (fig. 1, 3) and the declining importance of the library as a ‘gateway’ (fig. 7).

Utrecht has spoken about their experience at many events and published articles, for example:

Simone Kortekaas, Utrecht University. Thinking the unthinkable: A library without a catalogue – reconsidering the future of discovery tools for the Utrecht University Library [Kortekaas, S. 2012] – presented at the Liber Annual Conference, 2012

While this radical approach may not be for everyone, resource optimisation for web search engines is seen as a priority by many libraries and museums. One institution commented that although some librarians think that users just need to be trained better, the more likely explanation is that the library systems are not intuitive enough at present. At another, one third of access is through Google, one third through referral (eg links from VLE pages, portal search engine) and one third through direct access. Several participants commented that users are looking for a more Google-like experience but with more precise search results, something that is not easy to achieve.

There are different ways in which libraries optimise resource results for search engines, e.g.:

- Supplying record metadata to WorldCat to feed to Google
- Showing users that they can select their institution as their main library on Google/Google Scholar and navigate to the institutional e-database of choice
- Using Blacklight software

One institutional library noticed that although some users tended to start their search in Google, since the introduction of Primo, the number of article searches that started in Primo, rather than Google, had increased significantly.

Another university library said that while they ‘can’t fight the national behaviour of our users’, they highlight the Library website as the obvious option if users can’t find what they’re looking for via Google Scholar.

Several participants mentioned that one of the issues with optimisation for search engines was the extra effort required to keep the supplied data up to date. For example, resources stay in the index even when they no longer exist or have been superseded. Sometimes it is necessary to liaise directly with Google in order to have old links removed.

Another important issue is that users are not always clear that they are using a library resource if they start their search via a search engine, so providing clear library branding is an important consideration.

Clear explanations of how users can access resources off campus are important.

Other interesting information/ innovations

- Recent resource discovery events
    - Ivy League resource discovery day organised by Yale University, April 2015 http://campuspress.yale.edu/iviesplusdiscovery/
    - A one-day symposium at Harvard earlier this year that focused on the discovery of special collections material, January 2015 https://wiki.harvard.edu/confluence/display/SDI/Special+Collections+Discovery+Symposium
- Resource discovery set-ups that have inspired other institutions
    - Leicester University Library, UK (a tabbed approach)
    - Chalmers University of Technology, Sweden (a well presented search entry point)
    - University of Columbia, US (a customised home-grown Blacklight interface for multiple discovery tools)
    - Johns Hopkins University Library, US
    - Stanford University Library, US
    - A Civil war site created by one the US public libraries, an elegant and sophisticated search and resource access solution http://www.civilwaronthewesternborder.org/
    - Old NYC website (www.oldnyc.org), mapping historical photos from the NYPL, created by a volunteer, with the help of NYPL who supplied the records
- Some interesting examples of engaging users with the organisation's collections:
    - The Cleveland Museum of Art believes in providing a flexible and interactive experience for its users, embracing the users' close relationship with their mobile devices. They have developed Gallery One (http://www.clevelandart.org/gallery-one) which enables users to interact with the Museum's collections through their mobile devices.
    - The New York Public Library has been working on linking resources together in a meaningful way to help users make sense of their special collections and archives. They have also been engaging with popular online resources, such as Pinterest, to make their content (e.g. visual images) more discoverable on the web (www.pinterest.com/source/digitalcollections.nypl.org).
    - The British Library has developed several extensions for their Primo-based archive and special collection portal to present search results in a more semantic, linked way.
- In-house implementations:

The UK National Archives are a good example of a successful in-house implementation. However, most institutions don't have the same level of resource available to them. One

institution has developed in-house interfaces for their main discovery tools (e.g. OCLC Worldcat, SFX). Another, a joint national and university library, has developed separate discovery interfaces, one for the public and one for university users. Highly customised discovery solutions are better suited to the institution's specific needs but can be time consuming and costly to maintain.

User behaviour studies/ user feedback

Some useful user behaviour studies that were mentioned by the participants:

- **Ithaka Faculty Survey**, 2009 (mentioned earlier)
- **101 Innovations in Scholarly Communication** survey, ongoing (http://innoscholcomm.silk.co, https://innoscholcomm.typeform.com/to/Csvr7b/fallback)
    - Preliminary results (the first 1000 people) show that Google Scholar is the most popular research tool amongst the offered suggestions (92% of respondents use it; Please note that resource discovery platforms were not offered as a choice). Encouragingly, the majority of respondents use their institutional/ library access as their preferred tool/site for access to literature (93% of respondents selected that option)
    - For more information see https://101innovations.wordpress.com/tag/updates-insights/and also the presentation by B. Kramer and J. Bosman in June 2015 (http://www.slideshare.net/bmkramer/the-good-the-efficient-and-the-open-oai9, *The good, the efficient and the open - changing research workflows and the need to move from Open Access to Open Science* - OAI9, Bianca Kramer & J Bosman, June 2015
- **Report by Karen Calhoun** titled 'The Changing Nature of the Catalog and its Integration with Other Discovery Tools', 2006 (https://www.loc.gov/catdir/calhoun-report-final.pdf)The report and anecdotal evidence suggested looking for a discovery system that would provide users with a simpler, more Google-like experience, as more and more users were turning to search engines as the starting point of search (p. 26).

Future plans

Below is a list of some future improvements and developments that have been identified by the participants:

- Further integration of local and/or external holdings into the unified discovery interface
- Optimisation of resources for Google and other search engines
- Engaging with social media sites and other popular online resources (e.g. Pinterest)
- Exploration of linked data
- Improving the navigation aspect of a search, so that people can navigate their way through the results in a more joint-up semantic way, something that has already been given high priority on the web
- Streamlining off-campus campus

- Making access to full text easier (particularly from home when users are often presented with several options for access)
- Re-evaluating existing library education and user support to help users with their discovery needs, irrespective of where discovery starts
- Exploring research data
- Better use statistics to understand changing user behaviour
- Integration of all the different Museum systems together to pull reports using all the available data from multiple databases
- More 'born digital' content available to users

Conclusion

Resource discovery is an important area for organisations. Several participating institutions have resource discovery projects that are exploring similar issues to the University of Oxford. However, several participants felt that resource discovery is not always given as high organisational priority as other projects, e.g. digitisation. For all the participants resource discovery development is very much work in progress, as nobody felt that they had 'cracked it'. Resource discovery at a complex institution requires a lot of resource, especially if there is a lot of customisation and/ or local development required. This poses questions of long-term sustainability. Commercial discovery platforms make the job easier but there are limited options for customisation, user interface design and further development.

There is no commercial resource discovery system that has fully explored linked data, so search results tend to be linear and lack in detail. This is no longer sufficient for users who are getting used to a more semantic way of searching on the web. Sometimes not rushing to implement a new system that has come on the market pays off, as things move so fast. There is evidence of rapid improvements in functionality offered by commercial systems.

Optimisation of records for search engines and linked data are seen as important but are not fully explored at present. There are some good examples of innovative ways of linking content together, like the Old NYC example (www.oldnyc.org). Data science, automatic clustering, topic map generation, linking & grouping data in a meaningful way are some of the things that libraries need to start focussing on.

It is important to collect user behaviour statistics/ data in new ways, like watching Pinterest traffic of access to the organisation's collections, like in the visual images example from the New York Public Library.

The future of library discovery is about 'leveraging what happens outside the library and then providing materials to encourage discovery outside of the library'.

Areas of discussion with external institutions

Information to aim to gather before meeting/discussion

- Generic institutional information (e.g. size, departments, affiliated organisations etc)

- Library / Museums / IT structure (e.g. convergent service, library service structure)?
- Institutional information environment (e.g. library collections, museum & special collections, archives, research repositories, VLE etc)
- Current resource discovery mechanisms & products (including some or all of library, museums, archives and special collections, finding people/experts)

Discussion points (to be tailored to the specific points of interest for choosing the institution to consult)

- Background to current discovery services / products
  - Reasons for choosing current discovery tools
  - Important user requirements involved in decision / customisation (what would they say are the most important discovery needs of their users?)
  - Benefits and Gaps (the latter to include any important usability, accessibility issues encountered)
  - Important localisation work
- Current resource discovery structure
  - Tools for discovery of different material (integration of data from different local sources into combined discovery services, electronic journal article discovery, attitude to use of full text searching, etc)
  - Integration with the institutional website, other websites (e.g. VLE, college, faculty, museum websites)
  - User education - how users are encouraged to choose the right tool, discovery strategy etc (especially how to cope with differing discovery needs)
- Approaches taken to address different discovery routes used by patrons, e.g. via
  - the institutional website,
  - the library website/ portal,
  - museum website,
  - direct access to a resource or discovery via subject databases or publisher sites,
  - Google scholar/ Microsoft Academic Search/ other non-library routes
- Future plans (ask about motivation and use cases if possible)
  - Developments in existing tools
  - Use of new tools (e.g. social media)
  - Discovery for new data types (e.g. research data)
  - Collaborations & sharing (are they working with other organisations, or vendors; would they be interested in working with Oxford should their plans and ours coincide?)

Appendix 4:
Summary of Data from Vendors and Publishers

by Simon McLeish

Summary of Discussions (by Topic)

Discovery architectures

- Write once, expose in multiple formats
- Expose as XML (static web page possible through associated products)
- Expose through OAI-PMH
- The main thing is to help academics to burst out of [academic niche] boundaries. One way to do this is through “grey” search results – things that you need to know that you don't know you need to know.
- “Sharing information in a way that search engines can access will become a standard requirement in an ITT.”

Linked Open Data / RDF

A range of responses, including:

- Not yet looked at
- Interested in RDF, but not ready for use yet
- Some linked data work has been done, but not in UK project
- Beginning work on LOD
- Early days
- Important for linking between different types of data, such as people information to relate to discovered items associated with that person
- potential risk of vocabulary silos as data released with widely different structure etc
- Lodlam summit http://summit2015.lodlam.net/ working on LOD for libraries and museums
- schema.org used by major search engines & currently working on improvements to cover archive material

APIs and standards

- Harmonisation and standardisation work under way by some consultees
- Publishers keen to use KBART
- Response time is key when using APIs to extend searching

Bringing together distinct collections

- General interest from search providers in this area
- Some scepticism around difficulties of relevance ranking for disparate item types

Handling museum and archive data

- Beginning to be mentioned as a requirement for purchase of library discovery services
- Support under development
- Hierarchical data suggests browse-type solution
- Libraries have been working towards unification and standards for a long time. “To an extent, this is outside the library domain”, but it could be that this is the next step.

- Possibility of working on being able to upload Oxford University data to central index

Sending data to search engines

- Exposing local resources and outward discovery by users are not always linked (access issues)
- Importance of embedding discovery in an environment used by those you want to discover the items
- Concern at cleanliness and accuracy of data
- Ensure that the data they provide is fully utilised by the search engines

Discovering experts

- Featured results making it possible to promote existing research guides.
- Linked Open Data seems like a good fit here. Schema.org or bibframe descriptions of entities including people which have been identified in WorldCat. So the idea would be that if a search pulls in records related to an individual, or a place, or whatever, it would be possible to display further information and related links about the identified entity.

Social media/discovery

- Nothing ever happens by itself. The difficulty in academia is that all the communities are small and their members are in competition. This makes it harder to do discovery socially, so academics need to have help.
- They need to be able to find things non-socially too. Social discovery is not comprehensive enough, and those who rely on it will miss things; it doesn't suffice in the way it would have done 30 years ago.
- Usage of social media in resource discovery is generally as a starting point.

Common Themes

This consultation exercise is one in which it is harder to pull out common themes, because of the wildly different organisations consulted and because vendors have an agenda – they want to sell something or keep Oxford as a customer.

It is also worth mentioning that the Proquest purchase of Ex Libris since the consultation took place (and of which no advance warning was given by the consultees) may mean that some of what was said is no longer relevant – the likelihood of a convergence between Primo and Summon means that there are significantly fewer options for replacing Primo than there would have been earlier. For some analysis of the meaning of the purchase see (Grant 2015; Schonfeld 2015).

However, some common themes do still emerge.

1. There seems to be disagreement about the best course to follow for the future, both between participants in this strand and their thoughts and the outcomes from other strands. This is indicative of a lack of a consistent vision between different stakeholders, and/or the absence of any obvious, over-arching solution.
2. Work on Linked Open Data, which is seen by many as the closest there is to an over-arching solution, is something which is still in its early stages. Participants are not wholly convinced that the approach is without potential issues. But it may work well for discovery of people.
3. Sending data to search engines works best when the metadata is of high quality. On the other hand, providers of library search tools are interested in the integration of non-library (non-MARC) metadata.
4. Standards and APIs are important, and a continuing area of development.
5. There is scepticism about the use and value of social discovery.

Appendix 4:
Summary of Data from Vendors and Publishers

Bibliography

Grant, Carl. 2015. ‘Another Perspective on ProQuest Buying the Ex Libris Group.’ Blog. Thoughts from Carl Grant. October 8. http://thoughts.care-affiliates.com/2015/10/another-perspective-on-proquest-buying.html.

Schonfeld, Roger. 2015. ‘What Are the Larger Implications of ProQuest’s Acquisition of Ex Libris?’ Ithaka S+R. October 6. http://www.sr.ithaka.org/blog/what-are-the-larger-implications-of-proquests-acquisition-of-exlibris/.

Appendix 5: Literature Review 1: Understanding Resource Discovery

by Simon McLeish

Some general principles and descriptions of resource discovery are well documented in existing literature.

The Principle of Least Effort is described in (Bates 2003, p.48):

> People use the Principle of Least Effort, preferring easy-to-get information over harder-to-get information, no matter how high the quality of the latter, as a rule.

The same principle continues to be re-iterated in more recent analyses, including (Connaway, Dickey, and Radford 2011), and as part of Derek Law's analysis of the aliterate or post-literate world of the library of the future: “They want instant results and instant gratification because a fundamental tenet is that convenience trumps equality” (Law 2010).

As Bates goes on to say, though, this is mitigated by the impetus to put in greater effort “matters of great urgency or of great interest”, both factors which should apply to many resource discovery tasks carried out in the university sector.

In an earlier paper, Bates describes a more high level principle, which she calls “berrypicking” (Bates 1989), cited in (Falciani-White 2012). This describes an evolving strategy, refining how discovery is carried out as initial information discovered changes the conceptual model the user has of what they are looking for: “gathering information a piece at a time while the information need and search criteria continue to evolve” (Falciani-White 2012).

Janyk (2014) adds two more principles which are relevant to the methods used to carry out discovery tasks:

Rational Choice Theory suggests that individuals should be able to think through what they are trying to achieve and plan how best to reach their goals (e.g. by choosing multiple sources for information). Janyk suggests that this is not commonly seen in practice, as users of discovery services tend to blithely repeat methods which gave acceptable results in the past. This behaviour is more as predicted by *Gratification Theory*, where past success in finding relevant material means that the same method is re-used in the future. This analysis leads directly to the conclusion that the ways in which specialist discovery tools used in academia are approached will be similar to the ways in which the tools familiar to users before or outside the University have been successfully used, and unfortunately these methods often do not transfer well to the specialist tools.

A resource discovery process which is considered to be especially applied by undergraduates (and school students) is the Information Search Process described by Kulhthau (Kuhlthau 1988). This model divides the discovery process into task initiation, topic selection, prefocus exploration, focus formulation, information collection, search closure, and starting writing. Perceptions of anxiety decrease through this process, accompanying a progression from ambiguity to specificity, though others have observed that the process is not always followed in this order (Swain 1996). A more recent review

has confirmed that the model continued to be a valid description of the discovery methods of undergraduate students (Kuhlthau, Heinström, and Todd 2008).

Discovery Behaviour Specific to Undergraduates

Work has been done to analyse the unique aspects of the ways in which undergraduate students seek to obtain information – as reviewed in (Falciani-White 2012). The analysis presented there suggests that undergraduates carry out resource discovery principally to satisfy immediate academic requirements (essays, examinations, etc.), and is associated with a "certain amount of anxiety". As a result, convenience and familiarity outweigh suitability as criteria for services and methods used for discovery, and this is accompanied by "hesitancy" about asking for assistance from tutors or librarians – though it has always also been common for students to look for help and guidance from their peers.

Resource discovery methodologies also tend to evolve in sophistication – both in terms of what is to be discovered and also the approaches taken – with increasing experience. This has been been linked to the development in sophistication of undergraduates' epistemological beliefs, enabling them to evaluate the reliability, credibility and authority of sources, consider conflicting viewpoints, and use a wider range of discovery strategies (Whitmire 2004).

Discovery Behaviour Specific to Postgraduates

Most of the literature indicates that graduates would be expected to fall in between undergraduates and academic staff in terms of their resource discovery requirements and needs. They will be more experienced than undergraduates; they will be starting to use resources of types common in academic research but less so for undergraduate work (journal articles, datasets and so on). As summarised in (Falciani-White 2012), "graduate students could be said to fall along a continuum which features undergraduate (novice) information seekers at one extreme and faculty (expert) information seekers at the other".

Discovery Behaviour Specific to Senior Researchers

Differences in discovery behaviour between senior researchers and students are often described as a consequence of greater experience in discovery and research, along with more in depth knowledge of their specialist subject area. Ellis (1989), as cited by (Falciani-White 2012), described six typical behaviours seen in university faculty members:

- *starting*: reading reviews and review articles, initial exploratory searchers, etc. - actions to be undertaken before the main discovery exercise;
- *chaining*: tracking citations forwards and backwards from a known item;
- *browsing*: semi-directed searching, e.g. using author names or looking along a shelf of physical items;
- *differentiating*: using differences between items to determine relevance;
- *monitoring*: current awareness of activity in research field;
- *extracting*: systematic analysis of a specific source (e.g. publisher's web pages) to identify material of interest.

An individual discovery exercise may use some or all of the above, in any order (though usually iteratively building on prior actions). The discovery process of academic researchers can also be categorised as lying on a continuum between two extremes, with individuals using the methods deemed appropriate for the task in hand.

The Disciplines

Older literature on searching and discovery typically stereotyped the disciplines based on the understood research practices. At one end, the physical sciences and medicine are

characterised as those who are interested only in the most current information[18]. This group will generally perceive the academic publishing process as too slow, and will therefore focus on pre-print papers and circulated drafts, as provided by social media (or other informal circulation methods) and by services such as arXiv.

At the other end, the stereotypes associate the humanities with those who are interested in the full historical range of their subject areas. They are seen as heavy users of the libraries and museums, and interested in archives and other distinctive collections held locally (and be willing to travel across the world to access the right resource if necessary).

Recent research, though, shows "less overall difference between the physical sciences and the humanities" than expected (Meyer et al, 2011, p.18). A series of case studies across the disciplines in 2011 found that "from a broaders sociological view, it is striking how much consistency there is across the fields and disciplines" (Meyer et al, 2011. p. 18).

Context may also be important: different tools and methods may be used in work and home settings, even for the same tasks (Connaway, Dickey, and Radford 2011), but it is increasingly the case that people are trying different methods and tools and expect to do so (Dempsey 2007).

Bibliography

Arlitsch, Kenning, Patrick Obrien, Jason A. Clark, Scott W. H. Young, and Doralyn Rossmann. 2014. 'Demonstrating Library Value at Network Scale: Leveraging the Semantic Web With New Knowledge Work.' Journal of Library Administration 54 (5). Routledge: 413–25. doi:10.1080/01930826.2014.946778.

Asher, Andrew D, and Lynda M Duke. 2012. 'Searching for Answers: Student Research Behaviour at Illinois Wesleyan University.' In College Libraries and Student Culture: What We Now Know, edited by Andrew D Asher and Lynda M Duke, 71–85. Chicago, IL: American Library Association.

Bates, Marcia J. 1989. 'The Design of Browsing and Berrypicking Techniques for the Online Search Interface.' Online Information Review.

Bates, Marcia J. 2003. 'Task Force Recommendation 2.3 Research and Design Review: Improving User Access to Library Catalog and Portal Information: Final Report (Version 3).' Washington, D.C.: Library of Congress.

Bodleian Library. 1605. Catalogus Librorum Bibliothecae Publicae Quam Vir Ornatissimus Thomas Bodleius Eques Auratus in Academia Oxoniensi Nuper Instituit; : Continet Autem Libros Alphabeticè Dispositos Secundum Quatuor Facultates: Cum Quadruplici Elencho Expositorum S. Scriptur. Oxford: Apud Iosephum Barnesium.

Breeding, Marshall. 2015. 'The Future of Library Resource Discovery.' NISO White Paper. http://www.niso.org/apps/group_public/download.php/14487/future_library_resource_discovery.pdf.

[18] The 2011 report *Collaborative Yet Independent: Information Practices in the Physical Sciences (Meyer et al. 2011)* presents a varied picture of the information use of researchers in different areas of physics through a series of case studies, undermining the stereotype mentioned here.

Breitbach, William. 2012. 'Web-Scale Discovery: A Library of Babel?' … Discovery Tools in Academic …, 637–45. doi:10.4018/978-1-4666-1821-3.ch038.

Brindley, Lynne. 2010. 'Foreword.' In Envisioning Future Academic Library Services, edited by Sue McKnight, vii – ix. Facet Publishing.

Ciccone, Karen, and John Vickery. 2015. 'Summon, EBSCO Discovery Service, and Google Scholar: A Comparison of Search Performance Using User Queries.' Evidence Based Library and Information Practice 10 (1). University of Alberta: 34–49.

Cohen, Rachael A., and Angie Thorpe. 2015. 'Discovering User Behavior: Applying Usage Statistics to Shape Frontline Services.' The Serials Librarian 69 (1). Routledge: 29–46. doi:10.1080/0361526X.2015.1040194.

Colbron, Karen. 2015. 'Surf's Up – Observations from Recent Studies of Discovery.' Jisc Digitisation and Content Blog. http://digitisation.jiscinvolve.org/wp/2015/10/06/surfs-up-observations-from-recent-studies-of-discovery/.

Connaway, Lynn Sillipigni, Timothy J. Dickey, and Marie L. Radford. 2011. '"If It Is Too Inconvenient I"m Not Going after It:' Convenience as a Critical Factor in Information-Seeking Behaviors'. Library & Information Science Research 33 (3): 179–90. doi:10.1016/j.lisr.2010.12.002.

Dahlstrom, Eden, and Jacqueline Bichsel. 2014. 'ECAR Study of Undergraduate Students and Information Technology, 2014 - ERS1406.pdf.' Educause. http://net.educause.edu/ir/library/pdf/ss14/ERS1406.pdf.

Dempsey, Lorcan. 2005. 'In the Flow.' Lorcan Dempsey's Weblog. http://orweblog.oclc.org/archives/000688.html.

———. 2007. 'Discovery Happens Elsewhere.' Lorcan Dempsey's Weblog. http://orweblog.oclc.org/archives/001430.html.

Dooley, Jackie M., Rachel Beckett, Alison Cullingford, Katie Sambrook, Chris Sheppard, and Sue Worrall. 2015. 'Survey of Special Collections and Archives in the United Kingdom and Ireland.' In Making Archival and Special Collections More Accessible, 11–16.

Dooley, Jackie M., and Katherine Luce. 2015. 'Taking Our Pulse: The OCLC Research Survey of Special Collections and Archives.' In Making Archival and Special Collections More Accessible, 5–10. OCLC Research.

Ellis, David. 1989. 'A Behavioural Approach to Information Retrieval System Design.' Journal of Documentation 45 (3): 171–212. doi:10.1108/eb026843.

Falciani-White, Nancy. 2012. 'Understanding Information Seeking Behavior of Faculty and Students: A Review of the Literature.' In Planning and Implementing Resource Discovery Tools in Academic Libraries, edited by Mary Pagliero Popp and Diane Dallis, 1–21. IGI Global/Information Science Reference. doi:10.4018/978-1-4666-1821-3.ch001.

Favaro, Sharon, and Christopher Hoadley. 2014. 'The Changing Role of Digital Tools and Academic Libraries in Scholarly Workflows: A Review.' Nordic Journal of Information Literacy in Higher Education.

Flynn, Martin. 2010. 'From Dominance to Decline?: The Future of Bibliographic Discovery, Access and Delivery.' World Library and Information Congress: 76th IFLA General Conference and Assembly, 1–8. doi:10.1300/J111v27n03_f.

Freund, LeiLani, Christian Poehlmann, and Colleen Seale. 2012. 'From Metasearching to Discovery: The University of Florida Experience.' In Planning and Implementing Resource Discovery Tools in Academic Libraries, 22–43. IGI Global.

Frost, William. 2004. 'Do We Want or Need Metasearching?' Library Journal April 1: 27–30.

Green, Andrew. 2012. 'To What Degree Can You Now Use Digital Surrogates as a Preservation Strategy Now?' 'In Safe Hands? Guaranteeing Our Collections for Future Generations' RLUK/BLPAC Seminar. https://www.llgc.org.uk/fileadmin/fileadmin/docs_gwefan/amdanom_ni/dogfennaeth_gorffor aethol/darlithoedd_ac_erthyglau/dog_gorff_dar_erth_ish_12S.pdf.

———. 2015. 'The Future of Research Libraries.' The Anybook Bodleian Libraries Conference 2015.

Hall, Jeremy, and Michael Kucsak. 2014. 'Building the Next Successful Library Discovery Tool.' Library Faculty Presentations & Publications.

Jaggars, Damon E. 2012. 'Foreword.' In Planning and Implementing Resource Discovery Tools in Academic Libraries, edited by Mary Pagliero Popp and Diane Dallis, xiii – xiv.

Janyk, Roën. 2014. 'Augmenting Discovery Data and Analytics to Enhance Library Services.' Insights: The UKSG Journal 27 (3): 262–68. doi:10.1629/2048-7754.166.

Johnson, L., S. Adams Becker, V. Estrada, and A. Freeman. 2015a. NMC Horizon Report: 2015 Library Edition. Austin, Texas: The New Media Consortium.

———. 2015b. NMC Horizon Report: 2015 Museum Edition. Austin, Texas: The New Media Consortium.

Kay, David, and Owen Stephens. 2014. 'Improving Discoverability of Digitised Collections: Above-Campus and National Solutions.'

Kortekaas, Simone. 2012. 'Thinking the Unthinkable: A Library without a Catalogue -- Reconsidering the Future of Discovery Tools for Utrecht University Library.' LIBER: Re-Inventing the Library for the Future. http://www.libereurope.eu/blog/thinking-the-unthinkable-a-library-without-a-catalogue-reconsidering-the-future-of-discovery-to.

Kuhlthau, Carol C., Jannica Heinström, and Ross J. Todd. 2008. 'The "Information Search Process" Revisited: Is the Model Still Useful?' Information Research 13 (4): 45–45.

Kuhlthau, Carol Collier. 1988. 'Perceptions of the Information Search Process in Libraries: A Study of Changes from High School through College.' Information Processing & Management 24 (4): 419–27. doi:http://dx.doi.org/10.1016/0306-4573(88)90045-3.

Law, Derek. 2010. 'Waiting for the Barbarians: Seeking Solutions or Awaiting Answers?' In Envisioning Future Academic Library Services, edited by Sue McKnight, 1–13. Facet Publishing.

Luther, Judy, and Maureen C. Kelly. 2011. 'The Next Generation of Discovery.' http://lj.libraryjournal.com/2011/03/technology/the-next-generation-of-discovery/.

Meyer, Eric T, Monica Bulger, Avgousta Kyriakidou-Zacharoudiou, Lucy Power, Peter Williams, Will Venters, Melissa Terras, and Sally Wyatt. 2011. 'Collaborative yet Independent : Information Practices in the Physical Sciences.' Research Information Network RIN Report Series IOP Publishing 2011, no. december: 26. doi:10.2139/ssrn.1991753.

Michalko, James. 2015. ‘Foreword.’ In Making Archival and Special Collections More Accessible, 1–4. OCLC Research.

Morris Hargreaves Mcintyre. 2013. ‘Ashmolean Online Collections : Needs of Potential Users’, no. March.

Neal, James G., and Damon E. Jaggers. 2010. ‘Web 2.0: Redefining and Extending the Service Commitment of the Academic Library.’ In Envisioning Future Academic Library Services : Initiatives, Ideas and Challenges, edited by Sue McKnight, 55–68. Facet.

Nicholas, David, and Ian Rowlands. 2011. ‘Social Media Use in the Research Workflow.’ Information Services and Use. http://www.researchgate.net/profile/David_Nicholas5/publication/262272352_Social_media_use_in_the_research_workflow/links/00b49538357508798600000 0.pdf.

Noe, David Earl. 2012. ‘Replicating Top Users’ Searches in Summon and Google Scholar.’ In Planning and Implementing Resource Discovery Tools in Academic Libraries, 225–49. Information Science Reference.

OCLC Research. 2015. ‘Making Archival and Special Collections More Accessible.’

Petr, Julie, and Lea Currie. 2014. ‘How Do Librarians Prefer to Access Collections?’ In Proceedings of the Charleston Library Conference. doi:10.5703/1288284315598.

Plummer, Darryl, Leslie C. Fiering, Ken Dulaney, Mike McGuire, Claudio Da Rold, Adam Sarner, William Maurer, et al. 2014. ‘Top 10 Strategic Predictions for 2015 and Beyond: Digital Business Is Driving “Big Change”.’ https://www.gartner.com/doc/2864817?srcId=1-3132930041#a-152809420.

Poore, Megan. 2014. Studying and Researching with Social Media. SAGE Publications.

Poulter, Dale. 2012. ‘Primo Central: A Step Closer to Library Electronic Resource Discovery.’ In Planning and Implementing Resource Discovery Tools in Academic Libraries, edited by Mary Pagliero Popp and Diane Dallis, 535–43. IGI Global/Information Science Reference. doi:10.4018/978-1-4666-1821-3.ch031.

Priem, Jason, Heather a Piwowar, and Bradley M Hemminger. 2012. ‘Altmetrics in the Wild: Using Social Media to Explore Scholarly Impact.’ arXiv12034745v1 csDL 20 Mar 2012 1203.4745: 1–23.

Race, Tammera M. 2012. ‘Resource Discovery Tools: Supporting Serendipity.’ In Planning and Implementing Resource Discovery Tools in Academic Libraries, 139–52. Information Science Reference.

Ranganathan, Shiyali Ramamrita. 1931. The Five Laws of Library Science. Madras: The Madras Library Association.

Reidsma, Matthew. 2013. ‘Matthew Reidsma : The Library with a Thousand Databases.’ http://matthew.reidsrow.com/articles/58.

Schaffner, Jennifer. 2009. The Metadata Is the Interface of Archives and Special Collections , Synthesized from User Studies. OCLC Research. OCLC Research.

Schonfeld, Roger C. 2015. ‘Meeting Researchers Where They Start.’ Ithaka S+R. http://sr.ithaka.org/sites/default/files/files/SR_Issue_Brief_Meeting_Researchers_Where_They_Start_032615.pdf.

Seeliger, Frank. 2015. ‘A Tool for Systematic Visualization of Controlled Descriptors and Their Relation to Others as a Rich Context for a Discovery System.’ IFLA WLIC 2015 Cape Town. http://library.ifla.org/1227/1/141-seeliger-en.pdf.

Stevenson, Jane. 2015. ‘Exploring British Design: Research Paths.’ Archives Hub Blog. Accessed August 19. http://blog.archiveshub.ac.uk/2014/10/23/exploring-british-design-research-paths/.

Swain, D E. 1996. ‘Information Search Process Model: How Freshmen Begin Research.’ PROCEEDINGS OF THE ASIS ANNUAL MEETING 33. INFORMATION TODAY INC: 95–99.

Tang, Chris. 2015. ‘Achieving the Ables: Reimagining the Digital Discovery Services at the National Library of Singapore.’ IFLA WLIC 2015 Cape Town. http://library.ifla.org/1156/1/147-tang-en.pdf.

Tay, Aaron. 2015. ‘5 Things Google Scholar Does Better than Your Library Discovery Service.’ Musings about Librarianship. http://musingsaboutlibrarianship.blogspot.co.uk/2015/07/5-things-google-scholar-does-better.html#.VbiENpO37tT.

Walker, Jenny. 2015. ‘The NISO Open Discovery Initiative: Promoting Transparency in Discovery.’ Insights the UKSG Journal 28 (1): 85–90. doi:10.1629/uksg.186.

Webster, Peter. 2012. ‘The Web-Scale Discovery Environment and Changing Library Services and Processes.’ Planning and Implementing Resource Discovery Tools in Academic Libraries, 646–61. doi:10.4018/978-1-4666-1821-3.ch039.

Whitmire, Ethelene. 2004. ‘The Relationship between Undergraduates’ Epistemological Beliefs, Reflective Judgment, and Their Information-Seeking Behavior.’ Information Processing and Management 40 (1): 97–111.

Appendix 6:
Literature Review 2: A Survey of Technologies for Information Retrieval

by Alfie Abdul-Rahman and Saiful Khan, with closing remarks by Professor Min Chen

Introduction

This report summarises current technologies in library information-retrieval, search engine technologies, application of ontologies in search, learning of ontologies, search-related visualization, and search provenance. In this report, we have also carried out a user consultation examining the resource discovery tools that are currently used by researchers by interviewing both students and staff members of the University of Oxford. Part of the user consultation is also to explore and foresee what resources researchers will need in the future.

Resource discovery

In computer science, database management systems and information-retrieval study the computational methods for managing and discovering large amounts of data. Historically databases emerged from the requirements of accounting systems, e.g., e-commerce, banking, and reservations, and information- retrieval systems emerged from the requirements of library systems, e.g., books, bibliographic catalogs, and patent collections [WKRS09]. Therefore, regarding databases, research is mainly focused on developing different data models, improving consistency of search results, query processing optimisation, and efficiency. In information-retrieval, research is focused on the areas of text processing, ranking models, user satisfaction, and so on.

Databases

Databases that store structured data in a relational model (relations databases) are the most popular data management system. The need for preliminary blueprints and schema necessary for the relational data model made other alternative data models necessary. Moreover, presently structured, semi-structured, and unstructured data are being generated at unprecedented rates. As a result, there have been numerous attempts to create different data models, such as the document-oriented data model, graph data model, and key-value store which are among the most popular.

Document-orientated and graph data models address the problem faced by the relational data model, i.e., a future-proof complex schema. Such databases use a document-orientated data model that follows no internal structure, i.e., the fields and relations do not exist as predefined concepts. It acquires the type information from the data itself. All of the data attributes for an object are placed in a single entity (e.g., document or node) and are stored as a single entry. A relationship between the two entities is created using references.

Although the document-oriented and the graph data model eliminate the need for a future-proof complex schema to store and update data, the resources needed to enter records into a database are costly and expensive, particularly when the scale of the data is large. Resources are needed to transfer files into any database. Given the scale of files a large organisation or library has to manage with, the use of database does not appear to be a

feasible solution in terms of the resources that would be required to enter the files onto the database.

Information-retrieval

The area information-retrieval for managing and discovering information from a large collections of data originated from the discipline of librarianship. In early days, resources (e.g., books and manuscripts) were indexed using catalogs. Much like the index in the back of a book, the catalogs includes information about the resources. During search, the catalogs were browsed to find the location of the resources quickly. Such technique was first used by third century BC Greek poet Callimachus [ER07] and has always been the key feature of early catalog-based search, electromechanical machines, and modern computer-based information-retrieval systems.

Since 1891 to early 1950 there has been numerous work proposed on the development of electromechanical device assisting in rapid scanning of the catalog in library. These mechanical systems continued to be developed and used until the advancement of computer-based information-retrieval [Jah61].

In 1948, Holmstrom [Hol] described a computing machine called the *Univac* in a conference organised by The Royal Society, U.K. This system was capable of searching for text references associated with a subject code stored in a magnetic steel tape. This is to be known as the first computer-based system proposed for information search. A few years later, in 1952 an alternative approach was proposed by Taube *et al.* [TGW52]. They proposed the *Uniterm* system to index items by a list of keywords and such indexing technique is still being used today.

As information starts to explode, finding the relevant information among the long list of search results has become difficult. Therefore, ranking based on relevance has become essential. During 1970-1990, most of the fundamental ranking approaches (e.g., boolean, tf-idf, vector space, etc.) were developed, the details of such ranking techniques can be found in Manning's book [MRS09].

Apart from the discipline of librarianship, one major application of information- retrieval in real world was during 1990's. In late 1990, Berners-Lee created the World Wide Web and by 1993 the size of the Web become so large that Web-search has become essential. Web-search is one of the major application of information-retrieval in the real world. The previously used indexing mechanism remained the core component to efficiently store the crawled Web-pages along with the newly developed ranking algorithms such as HITS [Kle99] and *Google*'s PageRank [PBMW98] for ranking of Web-search results.

A detailed discussion on the evolution of information-retrieval domain starting from library to modern Web and enterprise search can be found in Sanderson and Croft's paper [SC12].

Another major application of information-retrieval was the Enterprise Search. With a growing amount of data in enterprises (e.g., libraries, compa- nies, etc.) managing and searching for information is becoming a challenge. The concept of Enterprise Search was introduced to encourage real-world application of information-retrieval [DSC10], where an enterprise develops an integrated data infrastructure, and enables its stakeholders to have an effective search interface for information retrieval.

Enterprise Search Engine

Numerous open source Enterprise Search Engines (e.g., Lucene, Solr, Elastic-Search, Xapian, LucidWorks, Indri, Terrier, and so on) were developed for text-based document

retrieval. They primarily use Lucene index at the backend. Lucene provides search index (also known as “inverted index”) creation, storage, and management facilities with document ranking algorithms (e.g., Boolean, TF-IDF, Cosine, Fuzzy, and so on). These modern tools also use probabilistic models to map documents to terms and then rank results. Probabilities are generated using methods such as TF-IDF and other language models. Although the exact methods of commercial systems are unpublished, it is likely they use techniques similar to Lucene and Indri.

Previous research in Enterprise Search Engine was focused on different mechanisms to improve the search time by enhancing the retrieval capability of Enterprise Search Engines for engineering documents [HLS15, HYLS14], multi-topic documents [LCH12], patents [MRL13], mechatronic data [ES11], and so on.

However, recent literature has focused mainly on improving the retrieval capability of an Enterprise Search Engine by knowledge-based support with ontologies for improved query enrichment and ranking, this is discussed further in the next section.

Ontology-based Retrieval

Ontologies have been widely used in domain-specific information retrieval. The Beagle++ [CCNP06] search engine was developed for Semantic Desktops. It indexes RDF metadata annotations to represent the meaning of documents together with the document content. The ranking of documents according to an ontology were reported by Kiryakov et al. [KPT+04] and Castells et al. [CFV07]. They presented an extension of the vector space model [SWY75] and together with document content, they indexed semantic annotations of documents and used this information for search.

All three promising approaches extended the vector space model using semantic information, however, none of them was able to apply measures of semantic association and have defined the annotations for the documents manually. Therefore, Billig et al. [BBL07] chose to align the representation of the document to the ontology automatically, using simple string matching.

Vallet et al. [VFC05] used ontologies to improve search over large document repositories. This retrieval model is based on an adaptation of the classic vector-space model, including an annotation weighting algorithm, and a ranking algorithm. Other work on file systems, e.g., [CGWX09, GSV07] enhanced the quality of file system search by leveraging the context of the search space. Castell et al. [CFV07] proposed a retrieval model based on an adaptation of the classic vector-space model, including an annotation weighting algorithm, and a ranking algorithm. This model built well- defined ontologies by populating knowledge bases and mapping keywords to concepts. Fernandez et al. [FCL+11] proposed an ontology-based information retrieval model by exploitation of domain knowledge bases to support semantic search capabilities in large document collection from a Web environment. ReFinder [DZW+13] is a context-based information re-finding system which assists users in re-finding their previously accessed files and web pages based on contexts such as time and location. A knowledge-based framework for integrating ontology-based personalised retrieval and reminiscence support was proposed by Shi and Setchi [SS13].

Leite and Ricarte [LR08] proposed multiple related ontologies for knowledge representation and integrated the system with a fuzzy model. A multi- ontology based multimedia annotation model was described by Dong and Li [DL06]. In their work, a domain independent multimedia ontology was integrated with domain ontologies to provide domain-specific views of multimedia content. Vaidurya [MS09] is a document search system developed for clinical document search by using a domain specific ontology. A semantic ontology-based framework for personal information management was proposed

by Xiao and Cruz [XC05]. Other query expansion techniques using ontologies can be found in works by Bhogal et al. [BMS07], Navigli and Velardi [NV03], and Díaz-Galiano et al. [DGMVUL09].

It is necessary to implement active learning process for updating ontologies dynamically as manually building such ontologies can be a tedious and time consuming task.

Ontology Learning

Ontology learning techniques have been widely studied. A comprehensive review and discussion of major issues, challenges, and opportunities in ontology learning can be found in Zhou's survey [Zho07]. Ichise *et al.* [Ich08] proposed a framework for organising the ontology mapping problem into a standard machine learning framework. This framework would use multiple concept-similarity measures. Maryam *et al.* [MSA11] described the bottleneck of ontology learning while dealing with the knowledge acquisition and modelling domain knowledge. They presented a survey of the different approaches in ontology learning from semi-structured and unstructured data. Idrissi *et al.* [EBB13] proposed a solution to overcome the manual knowledge acquisition bottleneck in ontology learning. They also presented a practical study of the methods that take data models as input to the learning process. Zhang *et al.* [ZHCZ13] proposed a semantic retrieval model of engineering domain knowledge, e.g., product design and process planning. They addressed the problems with existing keyword-based and semantic-enabled methods, and proposed an ontology-based semantic retrieval scheme for knowledge search and retrieval from domain documents.

The content-based retrieval systems have certain limitations. In a library, many of the files are non-textual (e.g., media, archive, images, scanned copies of invoices, books, catalogs, etc.). Therefore, improved metadata- based retrieval is essential. Similar to Web search, an Enterprise Search Engine can use relevance feedback information [DSC10] that can be used to learn and improve the ranking of the search results in the enterprise search [CDZ08] as no search can guarantee to find all the relevant file or how to correctly specify the search criteria. Most importantly, it would be useful to assist search with effective visualization and interaction [BYYG+08].

Library search

Gone are the days when the library was our first source for retrieving in- formation. Nowadays, a majority of students would either use *Google* (or *Google Scholars*) or other online search engines as their first choice for getting their information [GB05, DRCHW06]. The shifting towards *Google* and other online search engines as the point of information source could be due to frustration:

"Why is Google so easy and the library so hard?" [Ten09]

as voiced out by one of the participants during a focus group carried out by Tenopir in 2009. In an attempt to bring back users to the libraries, a huge number of higher education institutions have simplified the search process by implementing a single-search-screen that allows users to easily search all materials either owned by the institutions, accessible via subscription or open access using a single text box [Cau13, Hoe12]. This form of search is called *resource discovery* or simply known as the *library portal*. There are a number of surveys that have been carried out on resource discoveries in libraries and their implementations [Sto09, BCD14, SCOC13, Sto10]. For ex- ample, the implementation and evaluations of *Summon* at the University of Huddersfield [Sto10] and *Primo* at the University of Birmingham [BCD14].

Resource discoveries in libraries are powered at the back-end either using a *federated search* or *pre-harvested search* [Bac08, GGG09, Sto10, BYRN10]. In both federated search and pre-harvested search the users are allowed to search via a single access point. In federated search, the search word is looked for across multiple databases while in pre-harvested search the search word is looked for through the pre-harvested indexes of content with weighting

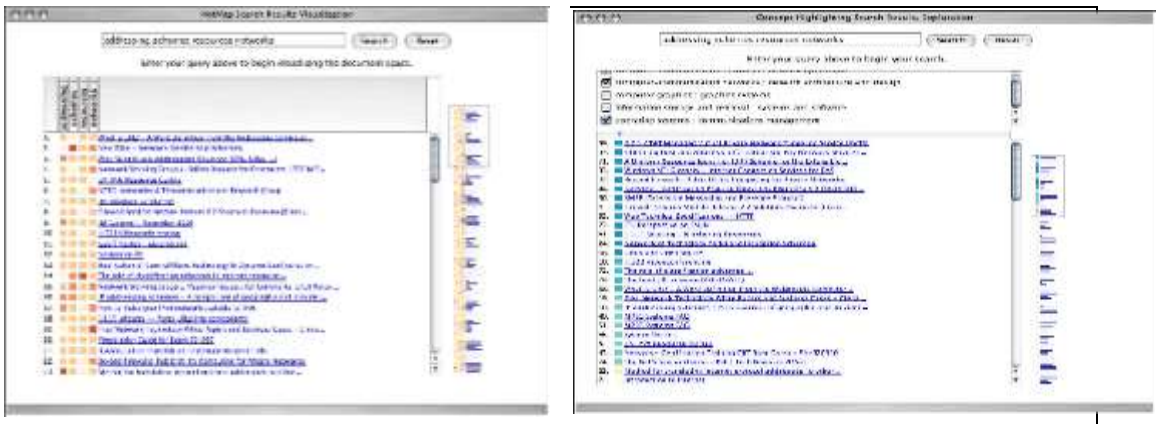


Figure 1: (left) HotMap: the term-frequency (tf) of each keyword in the query is shown using colour coding. (right) Concept Highlighter: Firstly a set of concepts related to the search query is displayed, and based on the user's selection the search results are sorted. The sorting is based on a fuzzy membership score and highlighted using colour coding. Adopted from [HY06].

applied to help with the relevancy ranking. By implementing the single-search-screen, users are able to retrieve results from multiple databases. For both of these searches, the results are then returned back often as a rank list in a paginated format. Although it is easier for users to retrieve search results using the single search box, users found that they have difficulties in comprehending the list of results that was returned to them [LSB13].

An alternative to a list of paginated results is a visualization approach that would allow users to explore and contrast the relationships and correlations of the retrieved results (or documents). For example, *Tag clouds* [Fei] and *inkblots* [AC07] that enable us to view the categorical information and statistical information about the documents, revealing various signature patterns for comparative analysis. Similarly, *ThemeRiver* [HHWN02] and *TIARA* [WLS+10] for visualizing changes of themes and topics over time within a collection of documents. For the visualization of multi-faceted relationships of keywords either within documents or across a large collection of documents *FacetAtlas* [CSL+10] could be used. Milne and Witten [MW11] presented *Ho¯ para* a visual search engine for exploring Wikipedia through its semantic relationship. Gove *et al.* [GDS+11] presented a visual analytics tool *Action Science Explorer* (ACE) for exploring academic publications through citations, ranking, and techniques of summarisation and automatic clustering.

Search-related Visualization

This section describes the research on the areas of search-related visualization mainly search result visualization and search provenance or history visualization. The work in the area of search-related visualization mainly focuses on the Web-search domain.

xFind [ASL+01] was developed to explore document search results as a ranked list as well as a scatter plot. In this tool, documents are displayed in a scatter plot where the search

attributes, e.g., relevance, file size, date, are mapped to *x*-axis and *y*-axis. The attributes can also be mapped to colour and the size of the icon. For example, the *y*-axis can correspond

Figure 2: Gradual refinement of query parameters in VisGets using the dimension: spatial – the region close to Caribbean is selected, (b) topical – the larger tag "politics" is selected, (c) and (d) temporal – first the month February is selected and then few days around the middle of February are brushed. Adopted from [DCCW08].

to the modification date of a document, and the document relevance can be mapped to icon size and/or icon colour (e.g., highly relevant in orange, and less relevant in white). Roberts *et al.* used glyphs to visualize web search results [RBR02] where the domains (e.g., .com, .edu, .net, .gov etc.) were mapped to the shape of a glyph.

In Web search, the search results are listed in 10–15 results per page and most users typically only scan the first page [SHMM99]. Such list-based representations limit the user to explore most of the search results. Therefore, the *HotMap* and *Concept Highlighter* were proposed by Hoeber and Xue [HY06] to support visual exploration of the Web search results, see Figure 1. These tools show search results in two levels of detail: an overview of the top 100 documents, and a detailed view of 20–25 documents at a time. The work also discussed how these views support the visual query of Web search results. Chau [Cha11] proposed a glyph based on a flower metaphor to visualize web search results.

Among all of the scholarly work that has been published by the visualization and information-retrieval communities on navigation and exploration of Web information space, the *VisGets* [DCCW08], *Fluid Views* [DCW12], and *PivotPaths* [DRRD12] are the most significant. *VisGets* [DCCW08] was developed to visualize different Web search attributes to assist in formulating search queries and to visualize search results using interactive geographic maps and tag clouds. Applications of *VisGets* to a variety of Web data collections were demonstrated by Dörk *et al.* [DWC09]. An empirical study on in- teractive and visual exploration of the Web using *VisGets* was also reported by Dörk *et al.* [DWC12], see Figure 2. *Fluid Views* [DCW12] enable users to view Web search results geographically, temporally, and content-wise in dual layers: for overview and details, see

Figure 3. *PivotPaths* [DRRD12] allows users to explore datasets with a variety of relation types across various dimensions, see Figure 4.

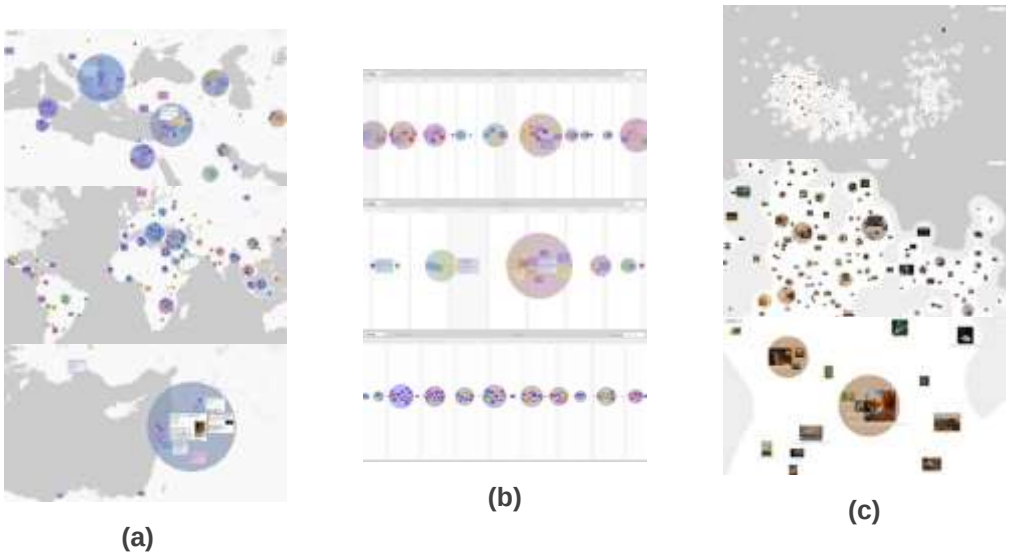


Figure 3: Fluid Views: (a) Zooming increases geographical detail in the map, here, from a continental view to the Mediterranean. (b) The time scale is changed from months, to weeks, and days. (c) Base layer gets gradu- ally magnified and the items, filtered, and displayed with more detail. Adopted from [DCW12].

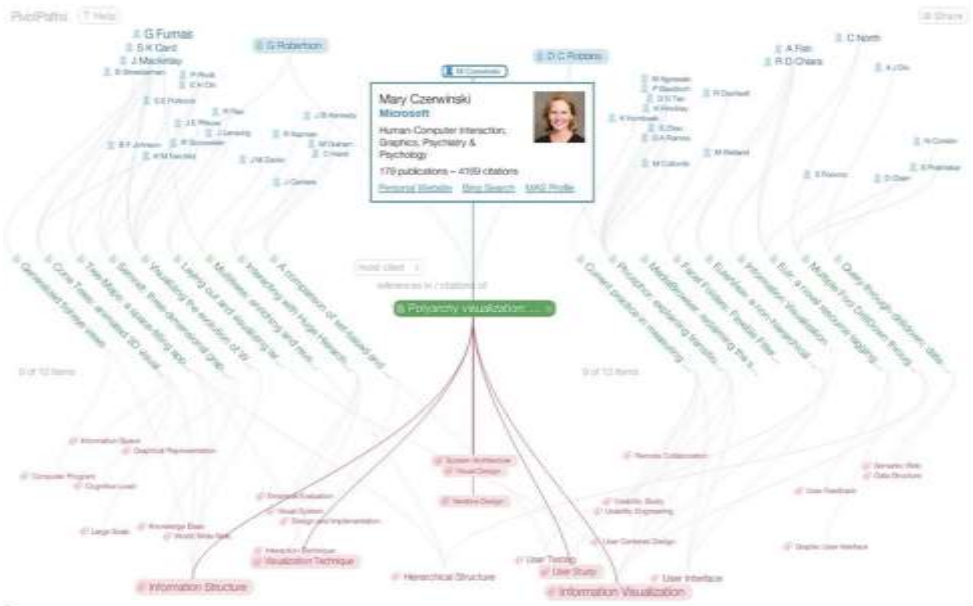


Figure 4: PivotPaths tool shows variety of relation types across various dimensions. Adopted from [DRRD12].

Another example of a notable work using Voronoi treemap to organise search results was proposed by Nocaj and Brandes [NB12], see Figure 5. ProjSnippet [GNSP+14] visualizes web search results in a global view and provides a clustering technique for identifying similar content, see Figure 6. The similarity measurement used in this work is based on the vector space model. The recent visualization supported tool for document search, Footprints [IDA+14], was developed for analysts searching for documents, Figure 7.

Tree-based Visualization for File System Search

The file systems resource discovery using visualization has been widely studied and considered by the visualization community. There are numerous file system visualization-assisted tools and techniques which have been

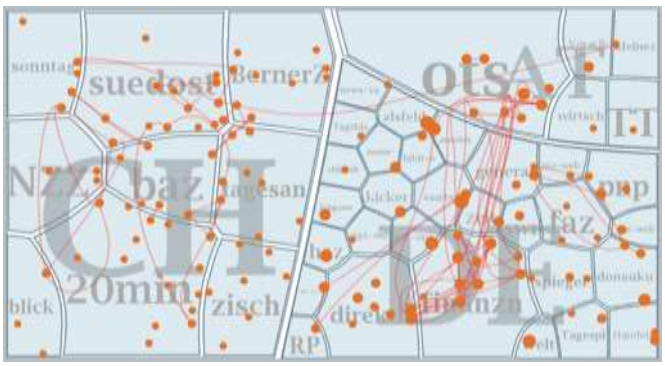


Figure 5: Search results are displayed with Voronoi treemap. Adopted from [NB12].

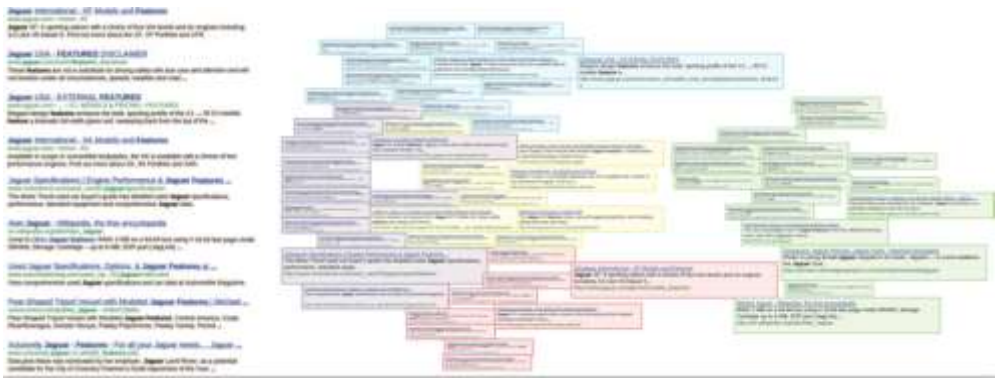
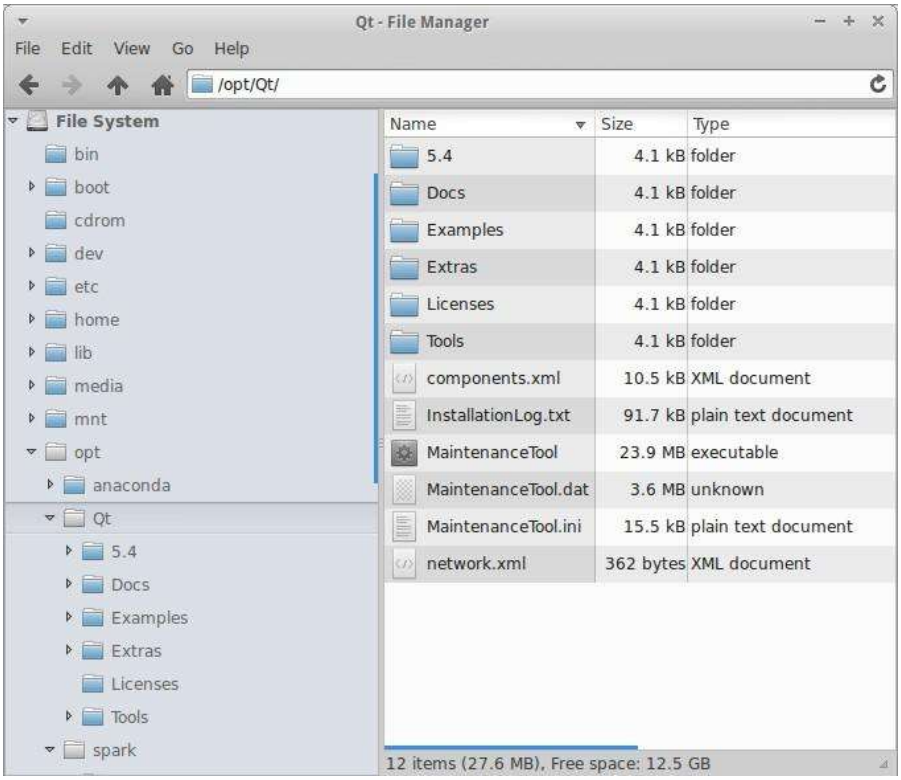
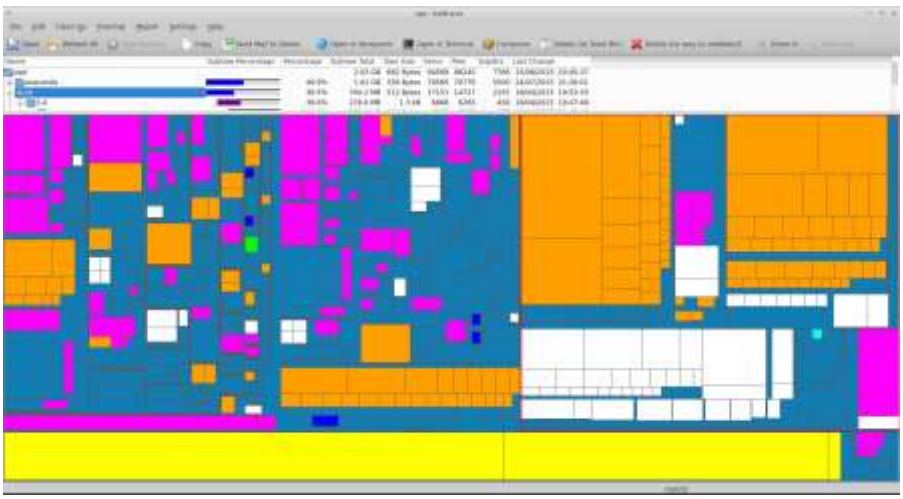


Figure 6: ProjSnippet tool visualizes the Web search result and clusters them based on content similarity. Adopted from [GNSP+14].

Figure 7: Footprints, a document search and visualization tool. Adopted from [IDA+14].



(a)

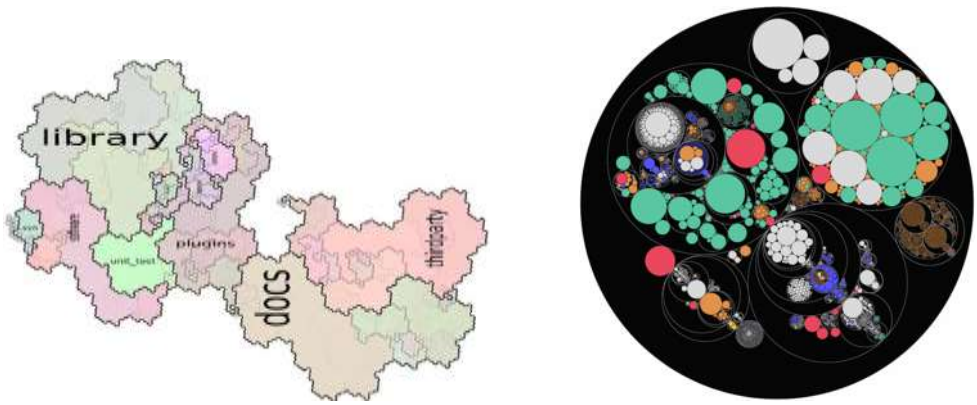


(b)

Figure 8: "/opt" directory of an example GNU/Linux system and its contents are visualized with (a) File System Explorer and (b) treemap (k4dirstat tool apps.ubuntu.com/cat/applications/k4dirstat).

developed for analysing disk usage, browsing hierarchy, searching for files and directories, and so on. The file system is a hierarchical structure, therefore, this essentially falls into the category of a hierarchy and/or tree visualization problem. A commonly used visualization technique for hierarchies is a tree or none-link diagram. For example, in all modern operating systems, the file system Explorer is the primary interface facilitating file system browsing and other functionalities. The Explorer displays the file system hierarchy using a tree as well as file/folder list/tabular view, see Figure 8 (a). Such explicit hierarchy visualization techniques do not utilise the space efficiently, therefore, fail to scale well for large file systems. Implicit hierarchy visualization techniques were proposed

Figure 9: Visualization of file search, featuring glyphs for search results (focus) and a treemap for the file system (context). Adopted from [KPW+14].



(a) (b)

Figure 10: Different treemaps with non-rectangular layout: (a) Circular (lip. sourceforge.net/ctreemap.html) and (b) Gosper [AHL+13].

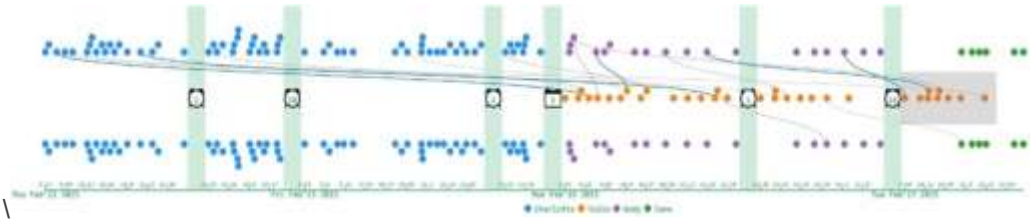
to solve this problem, a comprehensive survey on such visualization can be found by Schulz *et al.* [SHS11].

Treemaps were proposed by Johnson and Shneiderman in 1991 in their seminal paper [JS91, Shn92]. Their objective behind this work was to find out how and where the space of a hard drive is used by multiple users and to find large files that could potentially be deleted. In Figure 8 (c), the "/opt" folder was visualized using a treemap as an example.

Since the development of the first treemap, several improvements have proposed, e.g., performance (e.g., [SP]), layout (e.g., [Wat05, OS08, BDL05, Wet03, AHL+13]), texture (e.g., [VVdW99, LF08, BL07]), and so on. Figure 10 illustrates two non-rectangular tree-based layout.

Khan *et al.* [KPW+14] proposed focus+context visualization to visualize file search results of an Enterprise Search Engine. This work is the first serious attempt to introduce visualization in a large scale Enterprise Search Engine. In this visualization the search

results were shown as file system glyphs and the search space was visualized using a treemap, see Figure 9.



\

Figure 11: This is an example of a focus+context view of SPG displayed on a user's (Julie's) search provenance window. It illustrates how the user (focus) collaborated with other users (context). For example, Julie's fourth query is formulated based on Charlotte's while in Julie's second day of search she formulated two of her queries based on Andy's. The last nine queries from Julie's search are selected for glyph-based view as shown in Figure12. Adopted from [Kha15].

Such focus+context visualization assist users to quickly search the relevant results, and to identify false positives and false negatives.

Glyph-based Visualization of Search Results

Glyphs are visual entities composed of several visual channels representing multivariate qualitative and/or quantitative attributes. Search results are often displayed as text in a list or in a tabular view, often showing 10–15 results per page. However, most of the information e.g., search space, search attributes, search history, search results, their distribution, clustering pattern, and so on are not shown. Roberts *et al.* proposed glyphs to visualize Web search results [RBR02]. Chau [Cha11] proposed a flower metaphor based glyph to visualize more search attributes in a Web search results. Search keywords were shown using petals, outgoing links from a document were shown using leaves, the stem showed document length, and the supporting ground showed incoming links. Kachkaev *et al.* [KWD14] proposed glyphs with parallel coordinate plots to display and explore survey results. Karve and Gleicher [KG07] presented a glyph-based visualization for search results from the Protein Data Bank. Khan *et al.* [KPW+14] used glyphs for representing file system search result visualization, see Figure 9.

Provenance Visualization Related to Search

The historical information that needs to be captured over time to examine the quality or validity of data is called *Provenance Information*. In this section we shall discuss provenance visualization techniques applied to Web provenance data, e.g., search, navigation, click-stream, and query logs are extensively captured. Web search activities generate query logs [ZCV+12], usability logs [GBG96], and Web clickstreams [WSSM12]. *Clickstreams* [CPV+01, LPSH01] support the analysis of Web traffic and the users' Web path navigation behaviour [Chi02]. Such information is useful for understanding the effectiveness of Web marketing, advertising, and merchandising efforts. It also reveals how customers find stores, the products they browse as well as the products they purchase and so on.

Khan [Kha15] presented a comprehensive work on systematically storing and visualizing users' search provenance information where he introduced a graph structure called Search Provenance Graph (SPG) for recording the search provenance's data attributes and its structure. In addition, he developed multiple ontology-based knowledge supports for computing the

Figure 12: An example of glyph-based visualization of the queries selected in Figure 11. The search queries are shown above the central line, and the search results are shown beneath the central line. Each search query may feature four criteria of (in clockwise direction): keyword, type, size, and date. An icon appears when this criterion is new or has been changed. The search results are shown in two forms: summary statistics are shown when research results consist of more than 25 files and macro-glyph oth- erwise. For search results, green represents newly found files, grey rep- resent files common to the current and previous queries, blue represents files appeared only in the previous query, and red represents the selected files. The background colour of the 'Late Discovery' dialog is grey as there is currently no newly discovered file. Adopted from [Kha15].

semantic similarity between search queries by constructing and updating SPGs . In his thesis, Khan also proposed a combined glyph-graph visual representations for visualizing SPGs in a focus+context manner as well as three different types of visualization for representing the provenance information of a search related to a project:

- Focus+context view of SPG (Figure 11),
- Glyph-based view (Figure 12), and
- Collaboration Summary (Figure 13).

Such visualizations can help to assist users in (a) identifying the potential false positives in search results, (b) performing search tasks collaboratively, and (c) formulate/reformulate search queries based on previous searches

Challenges of resource discovery – user consultation

The purpose of the consultation is to examine what resource discovery tools are currently used by researchers as well as to foresee what resources they may need in the future. The interview began with a brief description explaining the objectives of the consultation followed by a demographics questionnaire. The interviewees were then asked a number of questions regard- ing their search tools and habits as well as what tools (or applications) they thought would further facilitate their research.

In total, ten people were interviewed. Of these interviewees, four were female (average age = 33) and six were male (average age = 39). The level of education varied from those who had completed a bachelors degree (1), to interviewees with a completed master degree or postgraduate diploma (3), to those who were either postgraduate students (2), or had

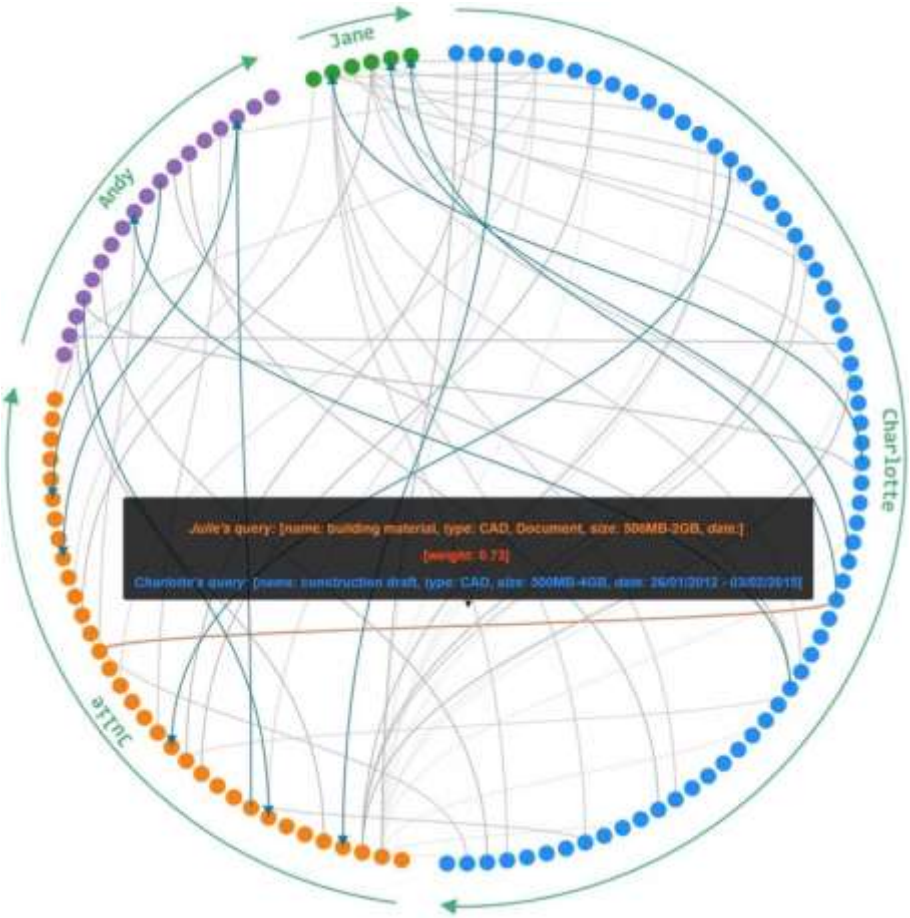


Figure 13: This example of collaboration summary view of SPG illustrates the similarity between the search queries performed by the administrators associated with a project. Adopted from [Kha15].

completed their Ph.D (4). All interviewees were recruited from the University of Oxford[8]. Eight interviewees were members of university staffs, and the other two were university students. The interviewees were from a number of disciplines, including computer science, humanities, physics, and engineering. All of the interviewees have completed at least a minimum of two projects that either involved research or required some form information gathering. A more detailed summary of the interviewees' background can be found in Table 1 search tools All of the interviewees were asked to describe how do they gather the information that they need for their research and a majority of them answered either *Google Scholars* or *Google*. *Google Scholars* is the preferred choice as it works well, and it requires low effort and minimal cognitive overload to use it. The interviewees also like *Google Scholars* as (i) it is quick, (ii) it provides good and targeted results, (iii) it is available at all times regardless of your location, (iv) it provides you with easy access to the full copy of the articles through the provided links of the published version, (v) it allows you to view the brief abstract providing you with the capability of judging the relevancy of the papers, (vi) it has the function to import citation into bibliography managers such as EndNote and BibTeX, (vii) it sends out a notification to indicate when there are new articles that might be of interest, and (viii) it allows you to view the authors and their list of publications, related articles, and the articles that have cited the article being viewed.

Qualification	Background	Area(s) of Research	Department	Job Role
PhD	Ancient History and Archeology, and Cultural Heritage	Digital Humanities, Linked Data	Oxford e-Research Centre	Research Associate
PhD	Electrical Engineering	Biomedical Imaging	Engineering Science	Professor
Masters	Computer Science	Visualization	Computer Science	Postgraduate Student
Postgraduate Diploma	Publishing and Fine Art	Digital Humanities	Oxford e-Research Centre	Project Manager
PhD	Computer Science	Computer Science and Social Science	Computer Science	Research Associate
Masters	Computer Science	Computer Science and Human Computer Interaction	Oxford e-Research Centre	Visiting Postgraduate Student
Masters	Egyptology	Egyptology	Pitt Rivers Museum	Administrator
Masters	Mathematics and Computer Science	Digital Humanities, Linked Data, Web System, and Semantic Web	Oxford e-Research Centre	Technical Lead
Bachelors	Mathematics and Computer Science	Web Linked Data	Oxford e-Research Centre	Research Software Engineer
PhD	Physics	Astrophysics	Oxford e-Research Centre	Research Associate

Table 1: A summary of the demographics information of the interviewees.

Despite its many advantages, there are also several drawbacks to Google Scholars mentioned by our interviewees for example (i) it returns back a vast number of results from which it is difficult to filter out the relevant articles, (ii) the detection of the omission of relevant articles becomes challenging due to the sheer volume of results and there is also concern that the list may not be accurate and up-to-date, (iii) as a keyword-based search, results can be skewed by the use of "incorrect" search terms, (iv) its inability to search across PDF documents for embedded scientific data, and (v) the lack of subject categorisations (or tagging) that makes it difficult to carry out a "top-down search" of drilling down a particular subject or retrieval of articles of similar (or associated) topic. There are also other drawbacks but they are more directed towards Google that (i) the results that are returned can sometimes be overwhelmed by commercial interests, e.g., adverts, and there is the question of the validity and providence of the information and whether it can be trusted. Semantic Scholars (https://www.semanticscholar.org/) has just been released and would not have been in the radar of our interviewees, but it might be worth for us to explore it.

The other answers on how do they gather the information that they need for their research are given below:

DOMAIN SPECIFIC DATABASES OR WEBSITES

SAO/NASA Astrophysics Data System for finding references and abstracts for astrophysics. The interviewee chose to use this database as it allows him to refine the search in a more precise manner (see Figure 14) as well as providing him with the capabilities to search for unpublished and pre-print articles for free.

Figure 14: SAO/NASA Astrophysics Data System interface that allows the users to refine their search in a more precise manner.

- *PubMed* for finding references and abstracts for life sciences and biomedical subjects.
- *Lexis*, *Nexis*, and *Westlaw* for retrieving legal and journalistic information and documents.
- *DBLP* for accessing journals and conferences proceedings for computer science.
- *Stack Overflow* for finding solutions to technical questions or computer programming problems.
- *Wikipedia* for finding general and basic information about a specific subject.
- *Pitt Rivers Museum Databases* for finding information about objects and historic photographs held in the Pitt River Museum.
- *Microsoft Academic Search* is an alternative bibliographic database for finding academic publications and was developed by Microsoft Research.
- *World Cat* for searching of items such as books and articles in a library that is near you.
- *IMSLP* for finding composers and music scores.
- *Web of Science* is an alternative bibliographic database for finding academic publications that is maintained by Thomson Reuters.

PEERS, COLLEAGUES, AND COLLABORATORS

- Consultation with collaborators on the articles and conferences to focus on. Project wikis are also a great source of information.
- Conversation with peers and colleagues on newly published articles and conferences to attend. These articles and conferences tend to be domain-specific.

SOCIAL NETWORKING SITES

- Through social networking sites such as *Twitter*. Authors, readers of articles, or attendees of conferences tend to tweet about papers that may be interest to their followers. These tweets are usually accompanied by a brief description as well as PDF links to the articles.

THESIS ADVISORS / SUPERVISORS

- Students are usually given an overview article by their thesis advisors as their starting point in information gathering.

DOMAIN SPECIFIC LIBRARIES

- *Pitt River Museum Library*, *Sackler Library*, and *Balfour Library* and colleges' libraries for specific information that is not available on-line.

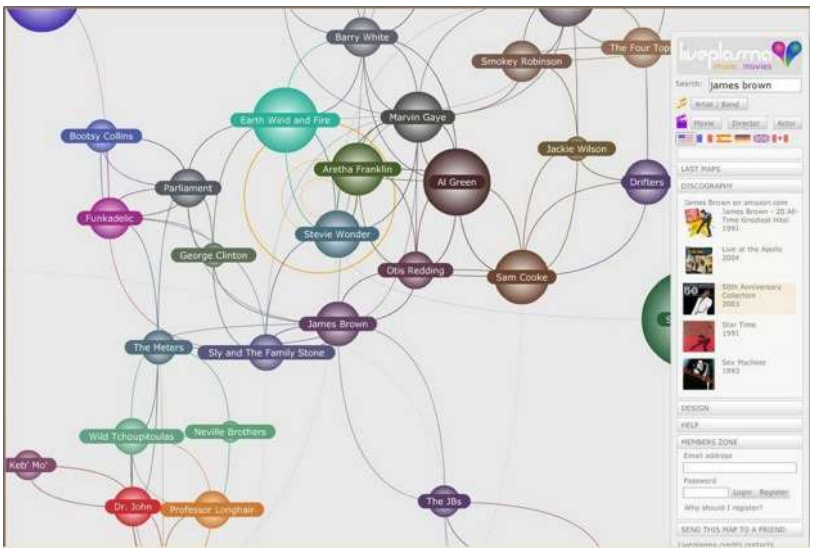
OTHERS

- **Oxford Talks**. A list of talks organised by the University of Oxford on domain-specific subjects.
- **Own Personal Books**. A majority of the interviewees have their own personal books that they refer to.
- **Through references in papers**. By manually building a timeline of how the topics and articles progresses across time using the references listed in the articles.

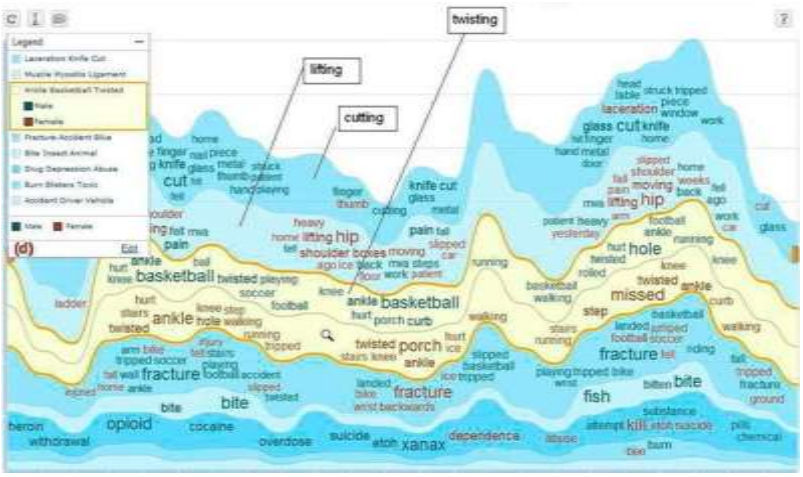
Search Terms

We then asked the interviewees to briefly explain on how do they come up with the search term. A number of interviewees explained that the search terms that they used are typically the results from conversations with colleagues, peers, and collaborators. Some of the search terms used are based on their experience from reading through articles as well as refinement of the search terms by heuristic. Library orientations were also mentioned as a source of recommendation for search terms as well as domain specific databases to search on.

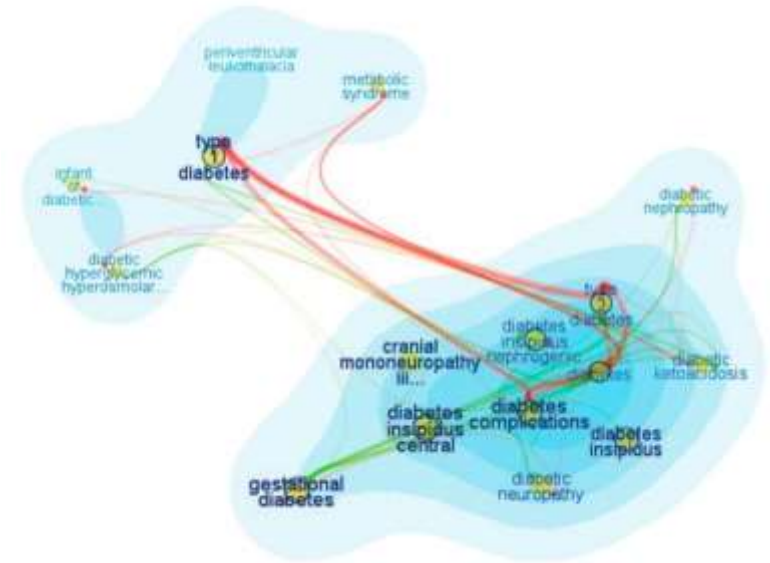
To get an overview of how familiar the interviewees were with the services that were provided by the Bodleian Libraries, we asked each of the interviewees to indicate their frequency in using each of the following catalogues and services:



(a)



(b)



(c)

Figure 15: Examples of visual search. (a) MusicPlasma [Pla]: A visual search engine for mapping music bands and artists. (b) TIARA [WLS+10]: A visual analytics system for exploring large collections of text by keywords across time. (c) FacetAtlas [CSL+10]: A visualization technique that allows you to view multi-faceted documents and keywords together with their relationship.s

SOLO (SEARCH OXFORD LIBRARIES)

- A majority of the interviewees answered never or very little of using SOLO. Only three interviewees answered: “Once a week or more”, “2-3 times a month”, and “Once a month”, respectively.
- We then asked the interviewees to briefly explain as to why SOLO was not being used. One of the interviewees explained that the up-to-date information that they require for their research are not available in books. (Although it is important to note that this may not be true for all subject areas.)
- Another reason why SOLO was not used was that SOLO does not provide you with the capabilities to search all available online bibliographic databases like *Google Scholars*.

CHINESE CATALOGUES

None of the interviewees have ever used the Chinese Catalogues.

ORA (OXFORD RESEARCH ARCHIVE)

At least three of the interviewees have occasionally used the ORA to check deposited data and theses.

SPECIAL COLLECTIONS CATALOGUES

None of the interviewees have ever used the Special Collections Catalogues.

Even though a majority of the interviewees do not use the four above mentioned services, they are happy with the collections of articles and publications that are made accessible online by the Bodleian Libraries.

information searched for

In order to get a better understanding of the interviewees' search habits, we asked the interviewees to briefly described on the type of information that they usually search for. We found that most of the interviewees usually search for the latest published articles and specialised information that are relevant to their fields. For science-based subjects, the interviewees would search for algorithms and experimental methods. Other information that was also searched for include patents, talks, overview explanations of specific topics, other researchers that are in a similar field, and who is doing what in specific subject areas.

visual search tool

Currently, the results that are returned by a majority of search engines and bibliographic databases appeared as a list in per page format. We asked our interviewees whether they have used a visual search tool before and if they are open to the possibility of using such tools in gathering information for their research. Figure 15 shows the visual search tools that were shown to the interviewees. Only one of our interviewees has used a visual search tool before. The visual search tool that the interviewee used in the past was for a linked map and he did not find the tool to be very useful. The interviewee is convinced that any future tools developed will struggle to match with *Google* as the tools created would not have access to the richness of data that *Google* has.

The other nine interviewees have not used a visual search tool before and were enthusiastic about the potential usage of a visual search tool. They felt that having such a visualization would be great for their research particularly if you are new to the subject area. Several concerns were also voiced with regard to using a visual search tool in research: the visualization is currently not highlighting what it not being displayed, and naive researchers may assume the results being displayed is a "definitive" guide and might not look explore further into the topic.

Appendix 6:
Literature Review 2: A Survey of Technologies for Information Retrieval

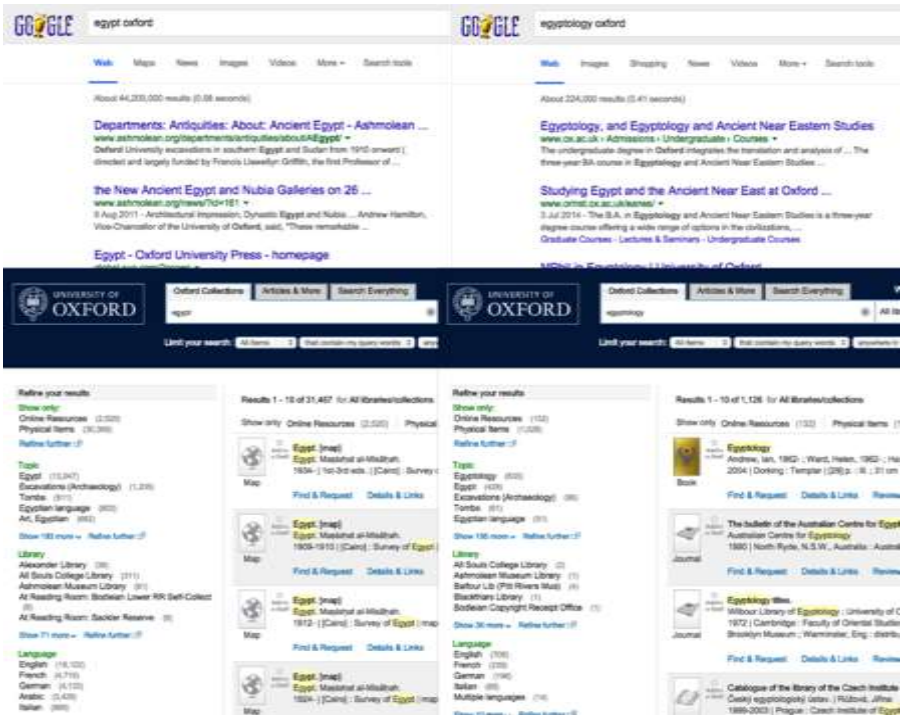


Figure 16: Web searches on Egypt and Egyptology in Oxford do not highlight all the possible resources that you can find in Oxford.

Wish list

We also asked our interviewees as to express what tools do they desire for the future and below are some of their answers:

A tool that allows you to:

- Generate the most common and related keywords and term-based subjects.
- Bridge the gaps between the different fields. Different fields could have different terms for the same concept, for example super-resolution is a term that is defined in imaging but in a different field it could be called something else. Until you have captured all of the “synonyms” of the term you could be missing out on a lot of references that may be important.
- Automatically generate the view for the co-citations of the articles as well as visualizing the degree of strength of the connections between the articles.
- Display the timeline of a specific topic and see how the topic popularity has increase or decrease over time.
- Visualize the timeline of the articles, i.e., the evolution of the paper by seeing who referenced it and who it referenced. There is, however, concern that such visualization can be too cluttered to view.
- Integrate between both the publications and social metadata where you would be able to receive social recommendation on articles and books as well as finding articles based on what the other researchers in similar fields are also accessing. The question of privacy is raised here as some subjects are so specialised that there are so few people in the field that you can easily determine who the researchers are.
- View the author’s profile and see their publications list.

- Search and extract embedded information inside PDF documents, such as embedded scientific data. For this, it may be useful for us to familiarise ourselves with the work that is currently carried out by the Visual Geometry Group at University of Oxford.
- An aggregator that would allow you to integrate between the different platforms, such as *Mendeley*, *Google Scholars*, and *JSTOR*.

Using the one of the visualizations, one of the interviewees pointed out that it would be nice to have a map of all the resources that are currently available in the University of Oxford. For example, if you are a new student in Oxford who has been assigned an essay to write on a specific topic in Egyptology, you might not have known that the Pitt River Museum, Somerville College, Sackler Library, the Ashmoleum Museum, etc. have collections of books and artefacts that may be of use to your research as a search on the web and SOLO would not have provided you with these resources information (see Figure 16). Some of the information such as the location of specific resources can only be obtained from talking with your peers, colleagues, or someone that has an extensive knowledge about the subject. It is important that such domain-knowledge is captured because if the source of information is no longer there then such information would be lost. When writing a prosopography or historiography article it is important for the writer to have access to collections of books and artefacts that they can trust and authenticate.

In one of our interview sessions, one interviewee made this comment: “For Humanities and Social Science scholars, repositories of books are important. However for Science-based scholars, it is less so as the information that they need are easily accessible online.” It is therefore important for the library to explore what online tools they can offer to these scholars that would facilitate their research. One suggestion is a tool that could help promote collaboration between researchers by enabling them to see the institutional affiliations of the researchers based on their expertise and allows you to set-up a chat system to communicate with these researchers. This tool can also be of potential use in facilitating the application of funding across multi-disciplinary fields by allowing scholars to see each other’s skills and expertise.

Instead of building another “search engine” for the library, perhaps we should explore the possibility of constructing a visual analytics tool that would complements the existing search engines and reference managers by implementing the items mentioned in the wish list by our interviewees and more. Perhaps this visual analytics tool can help bring back the “physical shelf browsing and serendipitous discovery” [LC14] that is slowly disappearing due to our online information seeking habit.

Concluding remarks

by Min Chen

From the above literature survey, we can make the following observations:

- Although the ontology-based search engine technology has been around for about two decades, it has not shown significant effectiveness in dis- covering library resources. The reasons that may explain this problem include the followings. (i) It is costly to capture and integrate metadata of library resources. (ii) It is costly to support reasonably complex ontologies with necessary techniques such as crawlers and indexing and buffering. (iii) The amount of search activities for library resources is simply insignificant for enabling ontology-learning.
- The database-based technology remains as the dominant search aids, but its deployment is hindered by the lack of support from visualization for enabling the

rapid identification of false positives and false negatives, from interactive visualization for exploring the search space without a set of well-defined search criteria, and from provenance visualization for reducing the cognitive load of remembering what has been searched.

- The next generation of technology for supporting the discovery of library resources may need to be developed through new innovations while learning the advances of other fields (e.g., online search). It is un- likely that a simply borrowing strategy will work. Such an uncreative strategy may actually damage library infrastructures and sciences in the long run, as the internet service providers are using the advantages of the online search technology to take away the services traditionally belong to library infrastructures. When there is a competition, it is disadvantageous for one party to compete on the other party's term and with the other party's technology.
- Interactive visualization and visual analytics should have a significant role in the next generation of library technology. It is important to understand what visualization is really for. The key is to save the users' time and reduce their cognitive load. However, most of visualizations used in the current library technology focus on displaying search results, while users often find easier and quicker to read the textual results anyway. Hence most existing effort was unproductive.

Oxford is one of the largest library resource providers in the world. It has the best opportunity to lead the development of library technology through innovation. However, it will always be harder to make and implement a strategic decision to innovate than to borrow.

references

[AC07] Ahmed Abbasi and Hsinchun Chen. Categorization and analysis of text in computer mediated communication archives using visualization. In *Proc. 7th ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 11–18, 2007.

[AHL+13] David Auber, Charles Huet, Antoine Lambert, Benjamin Re- noust, Arnaud Sallaberry, and Agnes Saulnier. GosperMap: using a Gosper Curve for laying out hierarchical data. *IEEE Trans. on Visualization & Comp. Graphics*, 19(11):1820–32, 2013.

[ASL+01] Keith Andrews, Vedran Sabol, Wilfried Lackner, Christian Gütl, and Josef Moser. Search Result Visualization with xFIND. In *Proc. 2nd Int. Workshop on User Interfaces to Data Intensive Systems*, page 50, 2001.

[Bac08] Murtha Baca. *Introduction to Metadata*. Getty Research Insti- tute, 2nd edition, 2008.

[BBL07] Andreas Billig, Eva Blomqvist, and Feiyu Lin. Semantic Matching Based on Enterprise Ontologies. In *On the Move to Meaningful Internet Systems*, volume 4803 of *LNCS*, pages 1161–1168. 2007.

[BCD14] S. Bull, E. Craft, and A. Dodds. Evaluation of a resource discovery service: Findit@bham. *New Review of Academic Librarianship*, 20:137 – 166, 05/2014 2014.

[BDL05] Michael Balzer, Oliver Deussen, and Claus Lewerentz. Voronoi Treemaps for the Visualization of Software Metrics. In *Proc. ACM Symp. on Software Visualization*, page 165, 2005.

Appendix 6:
Literature Review 2: A Survey of Technologies for Information Retrieval

[BL07] Renaud Blanch and Eric Lecolinet. Browsing Zoomable Treemaps: Structure-Aware Multi-Scale Navigation Techniques. *IEEE Trans. on Visualization & Comp. Graphics*, 13(6):1248–1253, 2007.

[BMS07]Jagdev Bhogal, Andy MacFarlane, and Peter Smith. A review of ontology based query expansion. *Information Processing & Management*, 43(4):866–886, 2007.

[BYRN10] Ricardo Baeza-Yates and Berthier Ribeiro-Neto. *Modern Information Retrieval: The Concepts and Technology behind Search*. Addison Wesley, 2nd edition, 2010.

[BYYG+08] Ori Ben-Yitzhak, Sivan Yogev, Nadav Golbandi, Nadav Har'El, Ronny Lempel, Andreas Neumann, Shila Ofek- Koifman, Dafna Sheinwald, Eugene Shekita, and Benjamin Sznajder. Beyond basic faceted search. In *Proc. WSDM*, page 33, 2008.

[Cau13] J. Caudwell. An A–Z of RDSs. *The Serials Librarian*, 65(1):1-24, 2013.

[CCNP06] Paul-Alexandru Chirita, Stefania Costache, Wolfgang Nejdl, and Raluca Paiu. Beagle++ : Semantically Enhanced Searching and Ranking on the Desktop. In *The Semantic Web: Research and Applications*, volume 4011 of *LNCS*, pages 348–362. 2006.

[CDZ08] Sara Cohen, Carmel Domshlak, and Naama Zwerdling. On Ranking Techniques for Desktop Search. *ACM Trans. on Information Systems*, 26(2):1–24, 2008.

[CFV07] Pablo Castells, Miriam Fernandez, and David Vallet. An Adaptation of the Vector-Space Model for Ontology-Based Information Retrieval. *IEEE Trans. on Knowledge and Data Engineering*, 19(2):261–272, 2007.

[CGWX09] Jidong Chen, Hang Guo, Wentao Wu, and Chunxin Xie. Search your memory ! - an associative memory based desk-top search system. In *Proc. SIGMOD*, pages 1099–1102, 2009.

[Cha11]Michael Chau. Visualizing web search results using glyphs: Design and evaluation of a flower metaphor. *ACM Trans. on Management Information Systems*, 2(1):1–27, 2011.

[Chi02] Ed H. Chi. Improving Web usability through visualization.*IEEE Internet Computing*, 6(2):64–71, 2002.

[CPV+01] Stuart K. Card, Peter Pirolli, Mija Van Der Wege, Julie B. Morrison, Robert W. Reeder, Pamela K. Schraedley, and Jenea Boshart. Information scent as a driver of web behavior graphs: results of a protocol analysis method for web usability. In *Proc. SIGCHI*, pages 498–505, mar 2001.

[CSL+10] Nan Cao, Jimeng Sun, Yu-Ru Lin, D. Gotz, Shixia Liu, and Huamin Qu. Facetatlas: Multifaceted visualization for rich text corpora. *IEEE Trans. Visualization and Computer Graphics*, 16(6):1172–1181, Nov 2010.

[DCCW08] Marian Dörk, Sheelagh Carpendale, Christopher Collins, and Carey Williamson. VisGets: coordinated visualizations for web-based information exploration and discovery. *IEEE Trans. on Visualization & Comp. Graphics*, 14(6):1205–12, 2008.

[DCW12] Marian Dörk, Sheelagh Carpendale, and Carey Williamson. Fluid Views: A Zoomable Search Environment. In *Proc. Int. Working Conf. on Advanced Visual Interfaces*, page 233, 2012.

[DGMVUL09] M C Díaz-Galiano, M T Martín-Valdivia, and L A Ureña- López. Query expansion with a medical ontology to improve a multimodal information retrieval system. *Computers in Biology and Medicine*, 39(4):396–403, 2009.

[DL06] Aijuan Dong and Honglin Li. Multi-ontology based multimedia annotation for domain-specific information retrieval. In *IEEE Int. Conf. on Sensor Networks, Ubiquitous, and Trustworthy Computing*, volume 2, pages 158–165, 2006.

[DRCHW06] C. De Rosa, J. Cantrell, J. Hawk, and A. Wilson. College students' perceptions of libraries and information resources: A Report to the OCLC Membership. Technical report, OCLC, 2006.

[DRRD12] M. Dork, Nathalie Henry Riche, G. Ramos, and S. Dumais. PivotPaths: Strolling through Faceted Information Spaces. *IEEE Trans. on Visualization & Comp. Graphics*, 18(12):2709– 2718, 2012.

[DSC10] Pavel Dmitriev, Pavel Serdyukov, and Sergey Chernov. Enterprise and Desktop Search. In *WWW*, pages 1345–1346, 2010.

[DWC09] Marian Dörk, Carey Williamson, and Sheelagh Carpendale. Towards Visual Web Search : Interactive Query Formulation and Search Result Visualization. *WWW Workshop on Web Search Result Summarization and Presentation*, pages 2–5, 2009.

[DWC12] Marian Dörk, Carey Williamson, and Sheelagh Carpendale. Navigating Tomorrow's Web: From Searching and Browsing to Visual Exploration. *ACM Trans. on the Web*, 6(3):1–28, 2012.

[DZW+13] Tangjian Deng, Liang Zhao, Hao Wang, Qingwei Liu, and Ling Feng. ReFinder: A Context-Based Information Refinding System. *IEEE Trans. on Knowledge and Data Engineering*, 25(9):2119–2132, 2013.

[EBB13] Bouchra El Idrissi, Salah Baina, and Karim Baina. Automatic generation of ontology from data models: A practical evaluation of existing approaches. In *IEEE 7th Int. Conf. on Research Challenges in Information Science*, pages 1–12, 2013.

[ER07] Simon Eliot and Jonathan Rose. *A Companion to the History of the Book*. Wiley-Blackwell, 2007.

[ES11] Oliver Eck and Dirk Schaefer. A semantic file system for integrated product data management. *Advanced Engineering Informatics*, 25(2):177–184, apr 2011.

[FCL+11] Miriam Fernández, Iván Cantador, Vanesa López, David Val- let, Pablo Castells, and Enrico Motta. Semantically enhanced Information Retrieval: an ontology-based approach. *Web Semantics*, 9(4):434–452, 2011.

[Fei] Jonathan Feinberg. http://www.wordle.net/. (Accessed on 2 Nov 2015).

[GB05] J. R. Griffiths and P. Brophy. Student searching behavior and the web: Use of academic resources and Google. *Library Trends*, 53(4):539 – 554, 2005.

[GBG96] M. Gray, A. Badre, and M. Guzdial. Visualizing usability log data. In *Proc. IEEE Symp. on Information Visualization*, pages 93–98, 1996.

[GDS+11] R. Gove, C. Dunne, B. Shneiderman, J. Klavans, and B. Dorr. Evaluating visual and statistical exploration of scientific literature networks. In *Visual Languages and Human-Centric Computing (VL/HCC), 2011 IEEE Symposium on*, pages 217–224, Sept 2011.

[GGG09] Ian Gibson, Lisa Goddard, and Shannon Gordon. One box to search them all: Implementing federated search at an academic library. *Library Hi Tech*, 27(1):118–133, 2009.

[GNSP+14] Erick Gomez-Nieto, Frizzi San Roman, Paulo Pagliosa, Wallace Casaca, Elias S Helou, Maria Cristina F de Oliveira, and Luis Gustavo Nonato. Similarity preserving snippet-based visualization of web search results. *IEEE Trans. Visualization & Comp. Graphics*, 20(3):457–70, 2014.

[GSV07] Karl Anders Gyllstrom, Craig Soules, and Alistair Veitch. Confluence: enhancing contextual desktop search. In *Proc. SIGIR*, page 717, 2007.

[HHWN02] S. Havre, E. Hetzler, P. Whitney, and L. Nowell. Themeriver: visualizing thematic changes in large document collections. *IEEE Trans. Visualization and Computer Graphics*, 8(1):9–20, Jan 2002.

[HLS15] Gyeong June Hahm, Jae Hyun Lee, and Hyo Won Suh. Semantic relation based personalized ranking approach for engineering document retrieval. *Advanced Engineering Informatics*, .(in press):., feb 2015.

[Hoe12] A. Hoeppner. The Ins and Outs of evaluating web-scale discovery services. *Computers in Libraries*, 32(3):6 – 10, 38 – 40, Apr 2012.

[Hol] J. E. Holmstrom. Section III. Opening plenary session. In *Proc. The Royal Society Scientific Information Conf.*, page 1948, London.

[HY06] O Hoeber and Xue Dong Yang. Exploring Web Search Results Using Coordinated Views. In *Int. Conf. on Coordinated and Multiple Views in Exploratory Visualization*, pages 3–13, 2006.

[HYLS14] Gyeong June Hahm, Mun Yong Yi, Jae Hyun Lee, and Hyo Won Suh. A personalized query expansion approach for engineering document retrieval. *Advanced Engineering Informatics*, 28(4):344–359, oct 2014.

[Ich08] Ryutaro Ichise. Machine Learning Approach for Ontology Mapping Using Multiple Concept Similarity Measures. In *7th IEEE/ACIS Int. Conf. on Comp. and Information Science*, pages 340–346, 2008.

[IDA+14] Ellen Isaacs, Kelly Domico, Shane Ahern, Eugene Bart, and Mudita Singhal. Footprints: A Visual Search Tool that Supports Discovery and Coverage Tracking. *IEEE Trans. on Visualization & Comp. Graphics*, 20(12):1793–1802, 2014.

[Jah61] Gerald Jahoda. Electronic searching. *The State of the Library Art*, 4:139–320, 1961.

[JS91] B Johnson and B Shneiderman. Tree-maps: a space-filling approach to the visualization of hierarchical information structures. In *IEEE Conf. on Visualization*, pages 284–291, 1991.

[KG07] Aneesh Karve and Michael Gleicher. Glyph-based Overviews of Large Datasets in Structural Bioinformatics. In *11th Int. Conf. on Information Visualization*, pages 1–6, 2007.

[Kha15] Saiful Khan. *Visualization Assisted Enterprise Search Engine*. PhD thesis, Department of Engineering Science, University of Oxford, 2015.

[Kle99] Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.

[KPT+04] Atanas Kiryakov, Borislav Popov, Ivan Terziev, Dimitar Manov, and Damyan Ognyanoff. Semantic annotation, indexing, and retrieval. *Web Semantics: Science, Services and Agents on the World Wide Web*, 2(1):49–79, 2004.

[KPW+14] Saiful Khan, Karl J Proctor, Simon Walton, René Bañares- Alcántara, and Min Chen. A Study on Glyph-based Visualization with Dense Visual Context. In *Comp. Graphics & Visual Computing*, pages 73–80. The Eurographics Association, 2014.

[KWD14] A. Kachkaev, J. Wood, and J. Dykes. Glyphs for Exploring Crowd-sourced Subjective Survey Classification. *Comp. Graphics Forum*, 33(3):311–320, 2014.

[LC14] Michael Levine-Clark. Access to everything: Building the future academic library collection. *portal: Libraries and the Academy*, 14(3):425–437, 2014.

[LCH12] Hsien-Tang Lin, Nai-Wen Chi, and Shang-Hsien Hsieh. A concept-based information retrieval approach for engineering domain-specific technical documents. *Advanced Engineering Informatics*, 26(2):349–360, apr 2012.

[LF08] Hao Lü and James Fogarty. Cascaded Treemaps: Examining the Visibility and Stability of Structure in Treemaps. In *Proc. Graphics Interface*, pages 259–266, 2008.

[LPSH01] Juhnyoung Lee, Mark Podlaseck, Edith Schonberg, and Robert Hoch. Visualization and Analysis of Clickstream Data of Online Stores for Understanding Web Merchandising. *Data Mining and Knowledge Discovery*, 5(1-2):59–84, jan 2001.

[LR08] Maria Angelica A Leite and Ivan L M Ricarte Ricarte. Fuzzy information retrieval model based on multiple related ontologies. In *20th IEEE Int. Conf. on Tools with Artificial Intelligence*, pages 309–316, 2008.

[LSB13] Cory Lown, Tito Sierra, and Josh Boyer. How users search the library from a single search box. *College & Research Libraries*, 74(3):227–241, 2013.

[MRL13] Tiziano Montecchi, Davide Russo, and Ying Liu. Searching in Cooperative Patent Classification: Comparison between keyword and concept-based search. *Advanced Engineering Informatics*, 27(3):335–345, aug 2013.

[MRS09] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schutze. *An Introduction to Information Retrieval*. Number c. Cambridge University Press, 2009.

[MS09] Robert Moskovitch and Yuval Shahar. Vaidurya: a multiple- ontology, concept-based, context-sensitive clinical-guideline search engine. *Journal of Biomedical Informatics*, 42(1):11–21, feb 2009.

[MSA11] Hazman Maryam, R. El-Beltagy Samhaa, and Rafea Ahmed. A Survey of Ontology Learning Approaches. *Int. Journal of Comp. Applications*, 22(9):36–43, 2011.

[MW11] David N Milne and Ian H Witten. A link-based visual search engine for wikipedia. In *Proc. 11th annual international ACM/IEEE joint conference on Digital Libraries*, pages 223–226, 2011.

[NB12] Arlind Nocaj and Ulrik Brandes. Organizing Search Results with a Reference Map. *IEEE Trans. on Visualization & Comp. Graphics*, 18(12):2546–2555, 2012.

[NV03] Roberto Navigli and Paola Velardi. An analysis of ontology- based query expansion strategies. In *Proc. 14th European Conf. on Machine Learning, Workshop on Adaptive Text Extraction and Mining*, pages 42–49, 2003.

[OS08] Krzysztof Onak and Anastasios Sidiropoulos. Circular partitions with applications to visualization and embeddings. In *Proc. 24th Annual Symp. on Computational Geometry (SCG '08)*, page 28, 2008.

[PBMW98] Larry Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The PageRank Citation Ranking: Bringing Order to the Web. Technical report, 1998.

[Pla] Music Plasma. Music Plasma. http://www.musicplasma. com/. (Accessed on 30 Oct 2015).

[RBR02] J Roberts, N Boukhelifa, and P Rodgers. Multiform glyph based web search result visualization. In *Int. Conf. on Information Visualization*, pages 549–554, 2002.

[SC12] Mark Sanderson and W. Bruce Croft. The History of Infor- mation Retrieval Research. In *Proc. of the IEEE*, volume 100, pages 1444–1451, 2012.

[SCOC13] V. Spezi, C. Creaser, A. O'Brien, and A. Conyers. Impact of library discovery technologies: A report for UKSG. Technical report, UKSG, Nov 2013.

[SHMM99] Craig Silverstein, Monika Henzinger, Hannes Marais, and Michael Moricz. Analysis of a Very Large AltaVista Query Log. In *Proc. SIGIR*, pages 6–12, 1999.

[Shn92] Ben Shneiderman. Tree Visualization with Tree-Maps - 2-D Space-Filling Approach. *ACM Trans. on Graphics*, 11(1):92–99, 1992.

[SHS11] Hans-Jörg Schulz, Steffen Hadlak, and Heidrun Schumann. The Design Space of Implicit Hierarchy Visualization: A Survey. *IEEE Trans. Visualization & Comp. Graphics*, 17(4):393–411, 2011.

[SP] Ben Shneiderman and Catherine Plaisant. Treemaps for space-constrained visualization of hierarchies. www.cs.umd.edu/hcil/treemap-history.

[SS13] Lei Shi and Rossitza Setchi. Ontology-based personalised retrieval in support of reminiscence. *Knowledge-Based Systems*, 45:47–61, 2013.

[Sto09] G Stone. Resource discovery. In H. M. Woodward and L Estelle, editors, *Digital Information: Order or anarchy*, pages 133–164. Facet Publishing, 2009.

[Sto10] G Stone. Searching life, the universe and everything? The implementation of summon at the university of huddersfield. In *Serials Solutions breakfast program at Internet Librarian International*, Oct 2010.

[SWY75] G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613– 620, 1975.

[Ten09] C Tenopir. Visualize the perfect search. *Library Journal*, 134(4):22, 2009.

[TGW52] Mortimer Taube, C. D. Gull, and Irma S. Wachtel. Unit terms in coordinate indexing. *American Documentation*, 3(4):213– 218, 1952.

[VFC05] David Vallet, Miriam Fernandez, and Pablo Castells. An Ontology-Based Information Retrieval Model. In *The Semantic Web: Research and Applications*, volume 3532 of *LNCS*, pages 455–470. Springer, 2005.

[VVdW99] J.J. J Van Wijk, H. Van de Wetering, and H de Wetering. Cush- ion treemaps: visualization of hierarchical information. In *IEEE Symp. on Information Visualization*, pages 73–78, 1999.

[Wat05] M. Wattenberg. A note on space-filling visualizations and space-filling curves. In *IEEE Symp. on Information Visualization (InfoVIS'05)*, pages 181–186, 2005.

[Wet03] Kai Wetzel. Pebbles—Using Circular Treemaps to Visualize Disk Usage, 2003.

[WKRS09] Gerhard Weikum, Gjergji Kasneci, Maya Ramanath, and Fabian Suchanek. Database and information-retrieval methods for knowledge discovery. *Communications of the ACM*, 52(4):56, 2009.

[WLS+10] Furu Wei, Shixia Liu, Yangqiu Song, Shimei Pan, Michelle X. Zhou, Weihong Qian, Lei Shi, Li Tan, and Qiang Zhang. Tiara: A visual exploratory text analytic system. In *Proc. 16th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, pages 153–162, 2010.

[WSSM12] Jishang Wei, Zeqian Shen, Neel Sundaresan, and Kwan-Liu Ma. Visual cluster exploration of web clickstream data. In *IEEE VAST*, pages 3–12, oct 2012.

[XC05] Huiyong Xiao and Isabel F Cruz. A Multi-Ontology Approach for Personal Information Management. In *Semantic Desktop Workshop*, 2005.

[ZCV+12] Jian Zhang, Chaomei Chen, Michael S. Vogeley, Danny Pan, Ani Thakar, and Jordan Raddick. SDSS Log Viewer: visual exploratory analysis of large-volume SQL log data. In *Proc. Visualization and Data Analysis*, volume 8294, page 13, 2012.

[ZHCZ13] Xutang Zhang, Xin Hou, Xiaofeng Chen, and Ting Zhuang. Ontology-based semantic retrieval for engineering domain knowledge. *Neurocomputing*, 116:382–391, sep 2013.

[Zho07] Lina Zhou. Ontology learning: state of the art and open issues. *Information Technology and Management*, 8(3):241–252, 2007.

Appendix 7:
Literature Review 3: Use of Social Media for Resource Discovery

by Simon McLeish

In 2015, social media, in all its forms, is ubiquitous. They can be categorised in various ways, but one frequently reproduced list of different types of social media is as follows (Nicholas and Rowlands 2011):

- Social networking.
- Blogging.
- Microblogging.
- Collaborative authoring.
- Social tagging and bookmarking.
- Scheduling and meeting tools.
- Conferencing.
- Image or video sharing.

There is already an extensive literature on the ways in which academics and students use social media in their work. The same paper records that in 2011, almost 80% of the surveyed academics were using social media in their research, a proportion rising to 95% in some subject areas. There is a correlation with youth, though not as significant a one as might be expected, and also the expected higher likelihood of use of social media by academics whose work predominantly requires cross-institutional collaboration. It is also not surprising that the different types of tool are considered to be most useful at different stages of the research lifecycle, some for identifying research opportunities or organising collaboration, others for disseminating research findings.

Poore (2014) (whose book should be consulted for a full length description of the use and potential of social media for both teaching/learning and research) gives the following list of the ways in which social media is used:

- Participation
- Collaboration
- Interactivity
- Community building
- Sharing
- Networking
- Creativity
- Distribution
- Flexibility
- Customisation

Of these ten roles, discovery has a part to play in at least five, either as part of the active participation or as a passive audience.

While it has been pointed out that much of the role of social media in academic endeavour is to facilitate existing practice rather than creating new working methods (Priem, Piwowar, and Hemminger 2012), others disagree (e.g. (Poore 2014)), and it seems likely that

research and teaching will both see the evolution of new methodology which takes advantage of social media.

One of the most important discovery-related applications of social media is to scholarly metrics. Today's scholars can take advantage of “altmetrics” both to measure the impact of their own work and as an aid for the discovery of well-regarded research articles, as discussed e.g. by (Priem, Piwowar, and Hemminger 2012). Altmetrics basically measure the informal citations of articles in various forms of social media, immediately giving a picture of the importance of an article which rounds out the information given by traditional citation counting (as well as being quicker to respond to new citations, and applicable to a wider set of academic outputs by including such things as research data). Recent work indicates that differences seen in earlier studies are being smoothed out as time passes (Colbron 2015).

Bibliography

Colbron, Karen. 2015. ‘Surf’s Up – Observations from Recent Studies of Discovery.’ *Jisc Digitisation and Content Blog*. http://digitisation.jiscinvolve.org/wp/2015/10/06/surfs-up-observations-from-recent-studies-of-discovery/ .

Nicholas, David, and Ian Rowlands. 2011. ‘Social Media Use in the Research Workflow.’ *Information Services and Use*. http://www.researchgate.net/profile/David_Nicholas5/publication/262272352_Social_media_use_in_the_research_workflow/links/00b4953835750879860000000.pdf .

Poore, Megan. 2014. *Studying and Researching with Social Media*. SAGE Publications.

Priem, Jason, Heather a Piwowar, and Bradley M Hemminger. 2012. ‘Altmetrics in the Wild: Using Social Media to Explore Scholarly Impact.’ *arXiv12034745v1 csDL 20 Mar 2012* 1203.4745: 1–23.