A Survey of Current Practices in Data Search Services

SiriJodha Khalsa (University of Colorado)
Peter Cotroneo (Elsevier)
Mingfang Wu (ANDS)

Report to RDA P11

1 Executive Summary

We examined the practices that data repositories employ in helping users search their holdings and demonstrate the value of looking across disciplines at the common data discovery issues, such as relevancy, that all data repositories face. We find that repositories desire guidelines for improving relevancy ranking in their systems, with small repositories having the greatest need. Repositories understand that their search systems need to be evaluated and improved, but often lack the resources (time and/or expertise) to explore and evaluate the available options. We conclude that there is an opportunity for people working in the search space to collaborate, through initiatives similar to TREC and the bioCADDIE 2016 Dataset Retrieval Challenge, to build test collections and other efforts that offer the greatest improvements in search services at the lowest cost.

2 Introduction

This work is carried out by the Relevancy Ranking Task Force of the Research Data Alliance (RDA) Data Discovery Paradigms Interest Group.

A critical element of any data repository is its search system, which determines the discoverability of a repository's holdings. The importance of this is reflected in the FORCE11 FAIR Guiding Principles for scientific data management and stewardship[1], which includes Principle F4: "metadata are registered or indexed in a searchable resource". The number of research data catalogues supporting this principle has accelerated over the past decade, e.g., re3data.org, the Registry of Research Data Repositories, had 23 repositories when it went online in 2012; the number quickly increased to over 1,200 data repositories from across the globe in three years (Pampel and Vierkant, 2015), and by September 2017, the registry had more than 1,900 repositories[2]. However, it is unknown how the catalogues of each have been configured and what the benchmarks are of the services used for making data searchable. Relevancy Ranking Task Force of the Research Data Alliance (RDA) Data Discovery Paradigms Interest Group developed a survey in order to gather information about what search systems are used, how they are configured and what has been done to improve data searchability. Our intent was that through analysis of the survey results we could:

- Help data repositories choose appropriate technologies when implementing or improving search functionality at their catalogues;

[1] http://www.nature.com/articles/sdata201618
[2] http://www.re3data.org/, accessed on 1st of August

- Provide a means for sharing experiences for improving relevancy ranking;
- Capture the aspirations, successes and challenges encountered by research data catalogue managers;
- Identify opportunity for collective action; and
- Serve as a benchmark for the future, to gauge how data search services have improved and matured.

3 Related Work

An information retrieval system consists of three basic parts: the *catalog* which contains representations of the resources of interest (documents, datasets, specimens, etc.), the representation of the user's information needs (the *query*), and a process for matching those needs with the resources described in the catalog, i.e., the *retrieval*. Index and retrieval are two key components of an information retrieval system. Index is to representing the resources that will be retrieved. Retrieval applies retrieval model to predict and explain what will be returned in response to a user's query (Goker and Davies,2009). There are three main retrieval models: exact match models, vector space models and probabilistic models. Exact match models look for exact matches between terms in a query and the contents of a catalog. Vector space models view documents and queries as high dimensional vectors, rank documents according to their distance (or similarity) to a query (Salton 1971), Probabilistic models rank "documents" by their estimated probability of relevance to a query, e.g. BM25 (Robertson and Spärck Jones 1976), language models that estimate the probability of generating the query from a document (Ponte and Croft 1998). Other models have been developed for special features and search tasks, for example, PageRank model that aims to retrieval high quality web pages from trusted sites by utilising link structure (Brin and Page 1998) among others. Each model has its strengths and weakness, thus selecting a model and turning parameters of a model for a particular collection and search tasks become experimental. Up to now, there is no guideline that says, 'If this is what your retrieval world is like, do this.' (Spärck Jones, 2006).

In parallel in developing models, the development of evaluation methods and methodology has played an important role in advancing research and development in the field of information retrieval. In general, evaluation of a search system follows Cranfield evaluation framework (Harman, 2013) that requires the creation of a test collection consisting three components: 1) a document collection to search on, 2) a set of queries that simulate user search tasks, 3) and relevance assessment of each document with respect to each query. A testing system is then fed with a set of queries and scored according to its ability to retrieval relevant documents, this is so called batch evaluation. An evaluation can also involve users to conduct search, mostly given a set of information search tasks, users' performance of completing search tasks with a testing system are then evaluated. Evaluations that involve users are usually costlier than the batch mode of evaluation. Furthermore, the batch and user evaluation methods may not give the same results. (Hersh et al. 2000, Wilkinson and Wu, 2004).

Constructing a testing collection is costly. In the early stages of the IR research, each research group built their own small test collection, such as the Cranfield test collection (Sparck Jones, 1981). As a result, it was difficult to compare systems across research groups and generalize the experimental results. Since the early 1990's, several initiatives began to address evaluation collectively as a community. Example initiatives include Text REtrieval Conference (TREC 1992-date)[3] (Harman 2005), Initiative for the Evaluation of XML Retrieval (INEX, 2002-2014)[4] (Fuhr et al. 2002), Information Access Evaluation meets Multilinguality and multimodality, and Interaction (2010-date)[5]. Each initiative addresses ever evolving resource types (e.g., from free text document, XML documents, images, web collections, to recently microblogs) and search tasks (e.g., ad hoc search, relevance feedback, novelty search, social book search, precision medicine and eHealth). These initiatives leverage the effort to build the evaluation platform - there is significant value for research teams working on the same search task using the same collection. What is evaluated by one team can be learnt by other teams, repeat work can be saved and collaboration can be merged as well.

In recent decades, we have seen a new type of resource emerging in academic repositories - data collections. There is an increasing number of public and domain specific data repositories appearing as discussed in Section 1. Enabling users to find data relevant to a research topic from either a repository or using a web search engine is a new challenge for search and data communities. In 2016, bioCADDIE announced a dataset retrieval challenge, focusing on "the retrieval of datasets from a collection that are relevant to the needs of biomedical researchers, in order to facilitate the reutilisation of collected data, and enable the replication of published results"[6]. This is the very first initiative from a data community to reach out to the search community to address the data search issue, and the challenge attracted 10 independent research groups who submitted 45 runs in total.

Overall, there has been little research and reports into what indexing methods and retrieval models are best suited to data repositories having structured metadata records representing data collections, and which may have links to publications, researchers, workflows and software. To understand how to better support data repositories in implementing current state-of-the-art search systems, we conducted a survey with the aim to understand where data repositories are currently and to establish a common ground upon which to progress.

4 A Survey of Data Search Systems

4.1 Instrument

Bearing in mind the goals as described in the Section 1, we designed the survey instrument to cover the following five survey question topics (SQT):

- **SQT1.** System configuration (e.g., ranking model, index methods, query methods) (7 questions)
- **SQT2.** Evaluation methods and benchmark (10 questions)
    a. What has been evaluated?

[3] Text REtrieval Conference (TREC): http://trec.nist.gov/

[4] INEX: Initiative for the Evaluation of XML retrieval: https://inex.mmci.uni-saarland.de/

[5] Conference and Labs of the Evaluation Forum: Information Access Evaluation meets Multilinguality, Multimodality, and Visualization: http://clef2017.clef-initiative.eu/

[6] bioCADDIE 2016 Dataset Retrieval Challenge: https://biocaddie.org/biocaddie-2016-dataset-retrieval-challenge

b. What evaluation methods have been applied?
c. How was the evaluation collection built?
d. What is the performance under specific configurations?

Note that these 10 questions were asked only if participants indicated that they had conducted some form of evaluation.

SQT3. Methods applied to boost searchability by web search engines (two questions)
SQT4. Other technologies or system configurations that have been employed (five questions)
SQT5. Wish list for future activities of the RDA relevance task force (two questions)

We also asked five additional questions about the characteristics of each repository so we could assess whether there is any correlation between repository characteristics and the adoption of a certain technology or configuration. Also, we included two additional questions about survey administration. In total the instrument asked 33 questions.

We performed two test runs with a small number of known repositories using a preliminary survey instrument, their feedbacks were accommodated in final version of the survey instrument. The final survey instrument and survey data are available on the research data repository Zenodo (https://zenodo.org/) (Khalsa, et al. 2018).

4.2 Survey participants

We targeted managers or developers of data repositories to participate in the survey. We compiled an initial list of 160 contacts from the Directory of Open Access Repositories, OpenDOAR (www.opendoar.org), from which we received 20 responses from an initial invitation. To acquire more participants, we then downloaded a list of 1,916 repositories from the URL https://www.re3data.org/api/v1/repositories (dated on: 24 July 2017), and used the re3data API (https://www.re3data.org/api/doc) to get each repository's contact information. We then filtered for repositories whose contact information included an email address, resulting in a list of 962 contacts. We then e-mailed the invitation to participate to those contacts, with a follow-up reminder several weeks later, and received more than 90 responses. A review of responses revealed 18 incomplete responses, i.e., a respondent had exited the survey without clicking the “Submit” button. Since many of these partial responses contained useful information we sent an email to those who had incomplete response explaining that their input was valuable and encouraging them to submit their responses even if they felt some of the questions did not apply to them, e.g., their systems did not do ranking at all.

In total, we received 99 completed responses; however, one of them provided very little information, so we did not include this response in the analysis. What follows is based on 98 completed responses.

5 Survey Result

This section summarises survey responses per survey question.

5.1 Repository Profile

Geographical distribution (Q1) The 98 responses represented repositories from three continents: Asia Pacific (14), North America (40) and Europe (44). Figure 1 shows a detailed distribution per country.

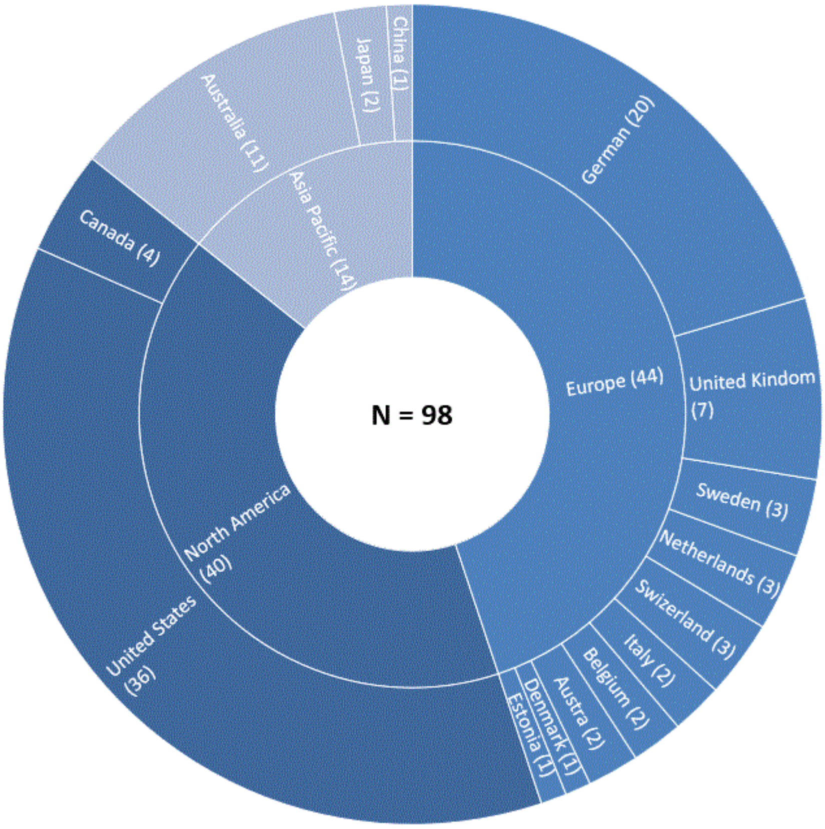

Figure 1. Geographical distribution of survey participants (N[7]=98)

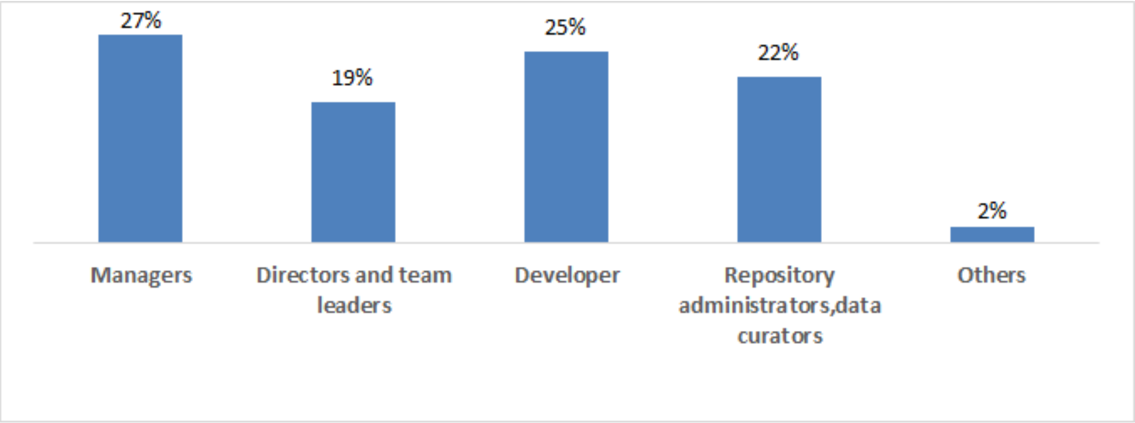

Figure 2. Distribution of respondents' role (N=91)

Respondents' role (Q2) Figure 2 shows the distribution of respondents' role. The manager role includes repository manager, project manager, data portal manager, and data manager.

[7] Notation: N is the actual number of responses to a survey question, so it varies across survey questions.

Apart from roles contained in the name, the director and team leader also includes principal investigators. The developer role includes software developers, architects, creator of repository, etc. Repository administrators include repository maintainers, administrators, data curators and coordinators of data registers. The “Others” category includes client support and postdoc.

Domain coverage (Q3) A multiple choice question asked what domains were covered by the repository. Figure 3 shows that 16% percent of the repositories indicated that they cover all domains. Twenty-eight percent selected more than one of the listed disciplines so in all 44% of the respondents support more than one discipline. A large number of repositories cover life science (35%), earth science (34%) and social science (28%); followed by physical science (17%), health science (12%) and computer science (3%). Among the 23% of responses that included “Other domains”, most of these responses were a further clarification of a chosen domain, for example, one response chosen “All Domains” but with a note “With a social sciences emphasis” in the “Other Domains” response. Three responses chose the domain “Social Science” but noted “linguistics” in the “Other domains”.

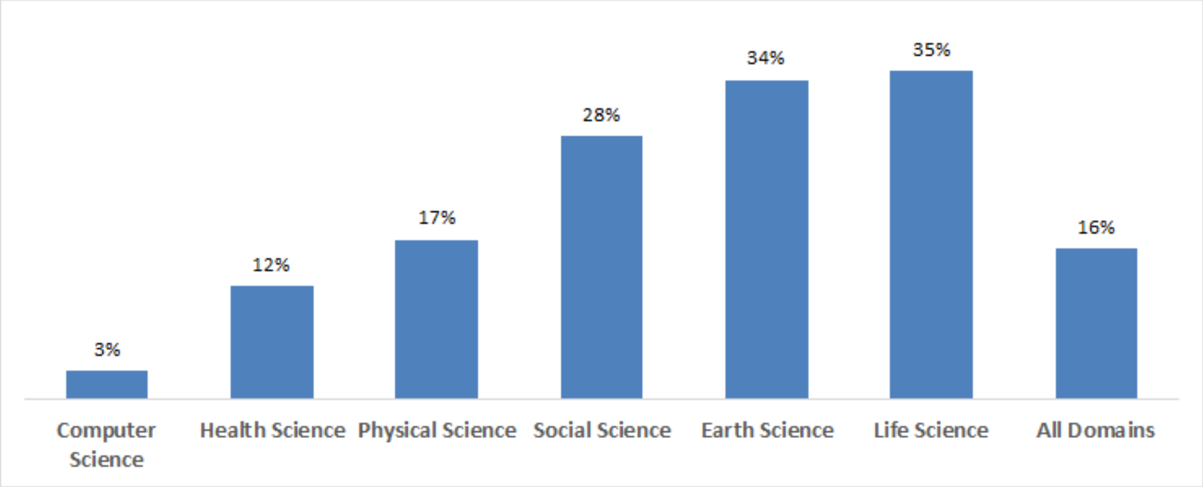

Figure 3: Domain Coverage (N = 98)

Data type (Q4) Figure 4 shows what objects have been registered in repositories. Ninety-seven percent of the responses chose “Datasets”. We determined that most of the respondents that did not choose the “Datasets” actually had datasets in scope, as noted by their “Other” responses, e.g. protein sequences and all related information, digitised special collection materials, and mass spectra. Apart from datasets, some repositories cover software (27%), published papers (24%) and methods (19%). One response chose Datasets, Published Papers and Thesis, but noted in the “Other” that “Taking only linked published papers and theses into account”, and another response noted in the “Other” that the repository includes “Data capture tools, e.g., survey, interview guides, etc.”.

Again, most of answers from the “Other” category are a further specification of types of datasets, e.g. audio, video, maps, scripts, Geographic Information System (GIS) data products, etc.

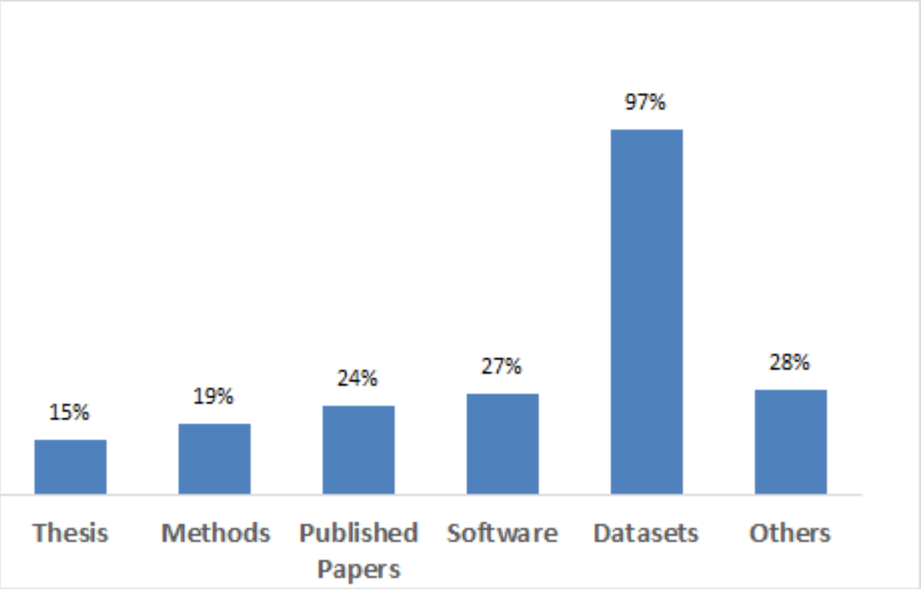

Figure 4. Data objects described by repositories (N = 98)

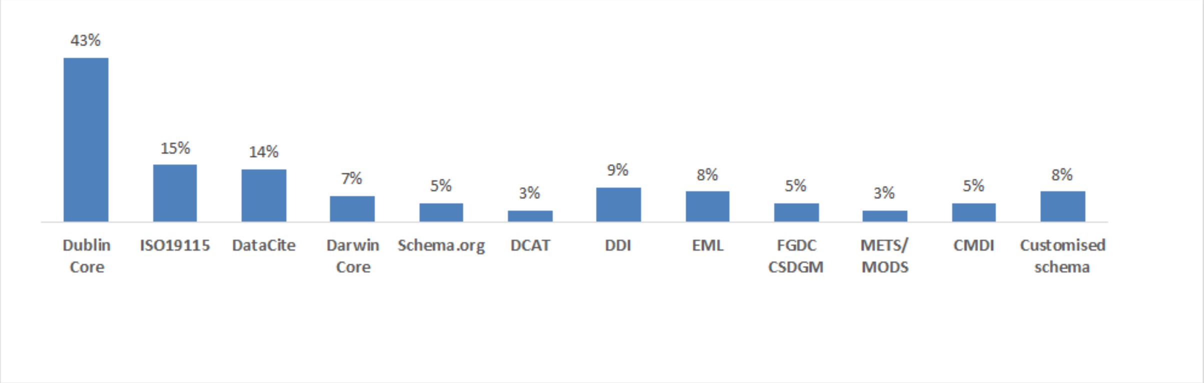

Figure 5. Metadata standards accommodated by repositories (N = 91)

Metadata schema (Q5) Respondents were asked to "List all metadata standards or schemas accommodated in your catalog". There were 91 responses in total to this question. As Figure 5 shows, the most accommodated standards or schemas are: Dublin Core (43%), ISO 19115 and variants (15%), DataCite (14%), DDI (9%), Darwin Core (7%), EML (8%), DCAT (3%), Schema.org (5%), and a few of customised ones (8%, without clarification which standards are customised from). Responses to this question also indicated that the majority of repositories cover more than one metadata schema, only 27% of responses listed a single schema.

Number of indexed records (Q6) On average, the repositories surveyed have about 50 million index records (the median number is 10,000), 17 repositories indexed more than 1 million records, eight repositories 100 million, and four repositories 500 million. When using browse method to discover datasets from a very large number of data records, search becomes an essential tool.

5.2 Search System Configuration

Search system (Q7) 96 out of 98 respondents answered the multiple choice question about what search systems are deployed by their repository. Figure 6 shows that most of them chose Apache Solr (31%), ElasticSearch (20%) and Apache Lucene (18%). Considering that both Solr and Elasticsearch use Lucene for search, when we fold the three answers together, search from a little bit more than half of repositories (54%) is powered by Lucene. The Lemur/Indri search system that has been used often by research groups was not picked up by any responded repositories.

There are 44% of responses choosing "Other or Out-of-Box Products", which includes; SQL/MySQL/MongoDB/PostgreSQL/PostGIS/Oracle/MariaDB (12), in-house implemented system (8), explanation (11, some are related to their search systems, some have nothing to do with search system), and Others (e.g., DSpace, Eprint, Figshare, etc).

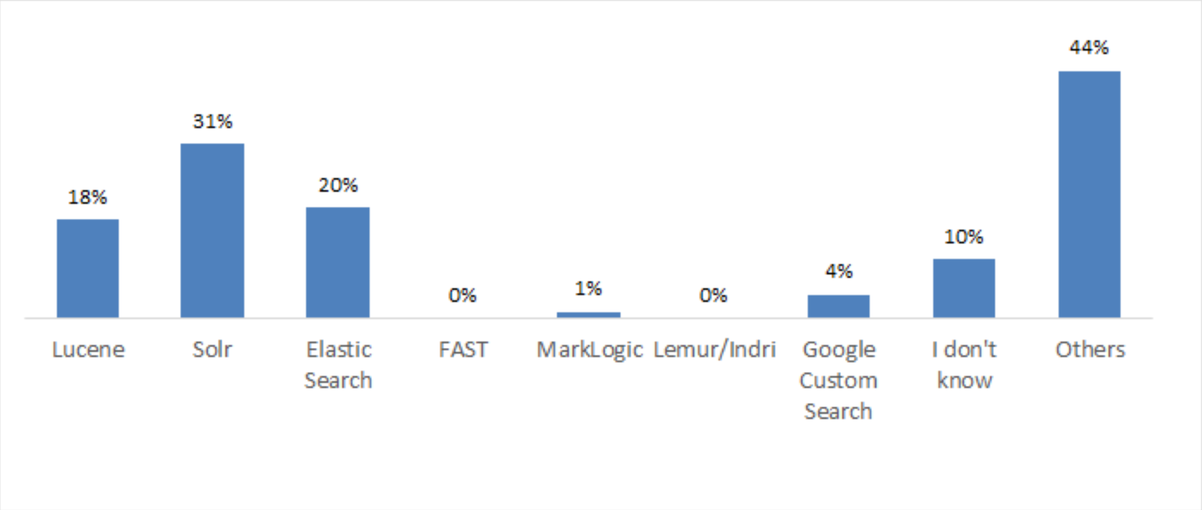

Figure 6. Search Systems (N = 96)

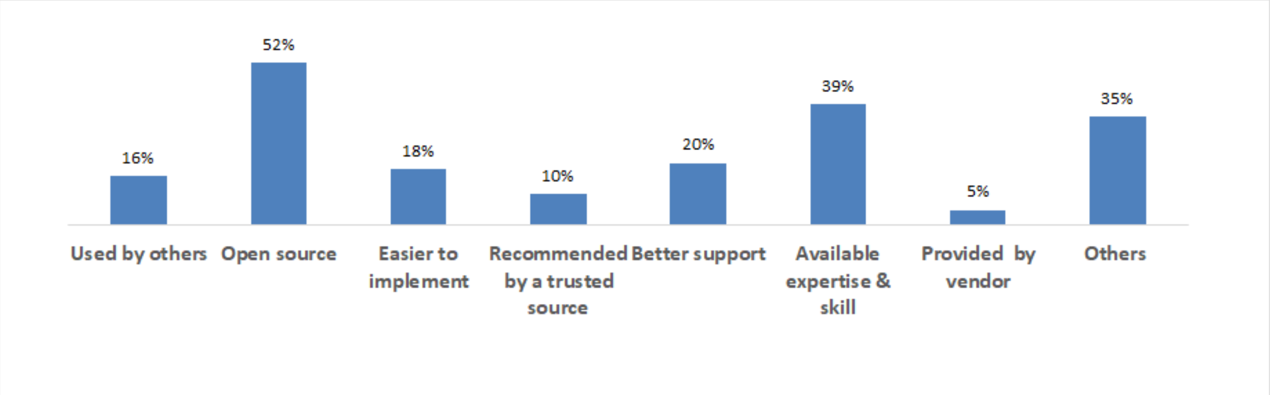

Figure 7. Reasons for choosing a search system (N = 96)

Reasons to choose a search system(Q8) Figure 7 shows what were the reasons for choosing current search system. Open source (52%) and available expertise and skill (39%) are the top two reasons, followed by having better support (e.g. documentation and community support), easier to implement (18%) and used by others (16%). "Others" reasons include “performance with large numbers of records”, “It was the request of the scientific community”, “minimal administration”, and “implementation flexibility”.

Indexed content (Q9) This question concerns what was included in their search index. Figure 8 shows most of repositories indexed either all available metadata (58%) or partial metadata (52%). Note there are 14 overlaps among 94 responses that chose both partial metadata all available metadata; 9 out of 14 overlaps have a number of standards and schemas (Q5), it is possible that each schema is treated differently. Data dictionary/variables (33%), metadata from publications (29%), text from publications, and tags from users (20%) are all indexed by a number of repositories. In the “Other” category were “peak data”, “spatial indexing”, “administrator tags”.

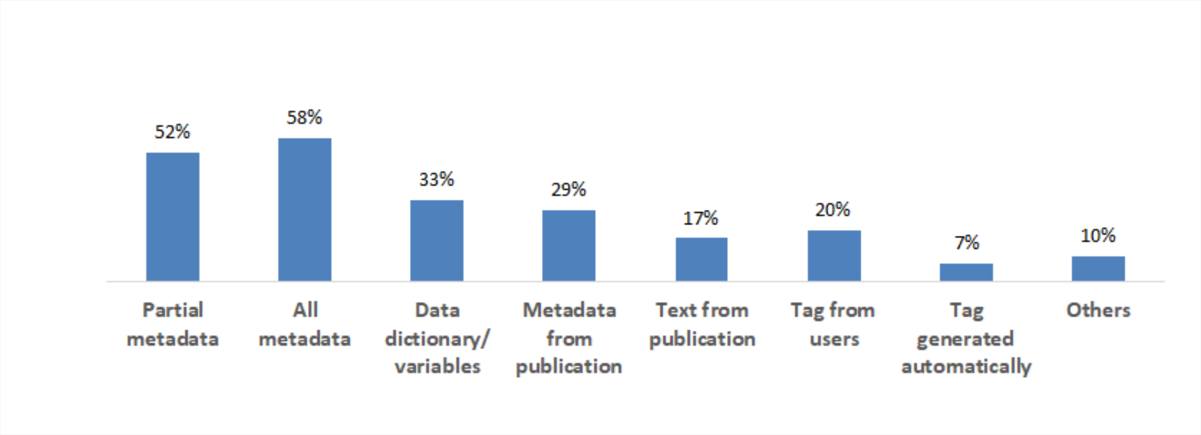

Figure 8. Indexed elements (N = 98)

Supported query construction strategies (Q10) In terms of what query construction strategies, majority of repositories offer free-text query (79%). Repositories also offer other querying options (cf. Figure 9): advanced query, e.g. Boolean query, phrase query, and proximity query etc. (50%), faceted query (45%), spatial query (44%), temporal query (44%). Out of those 52 responses that support spatial and/or temporal queries, they tend to support both query types (65%). Not surprisingly, spatial or temporal query support mostly comes

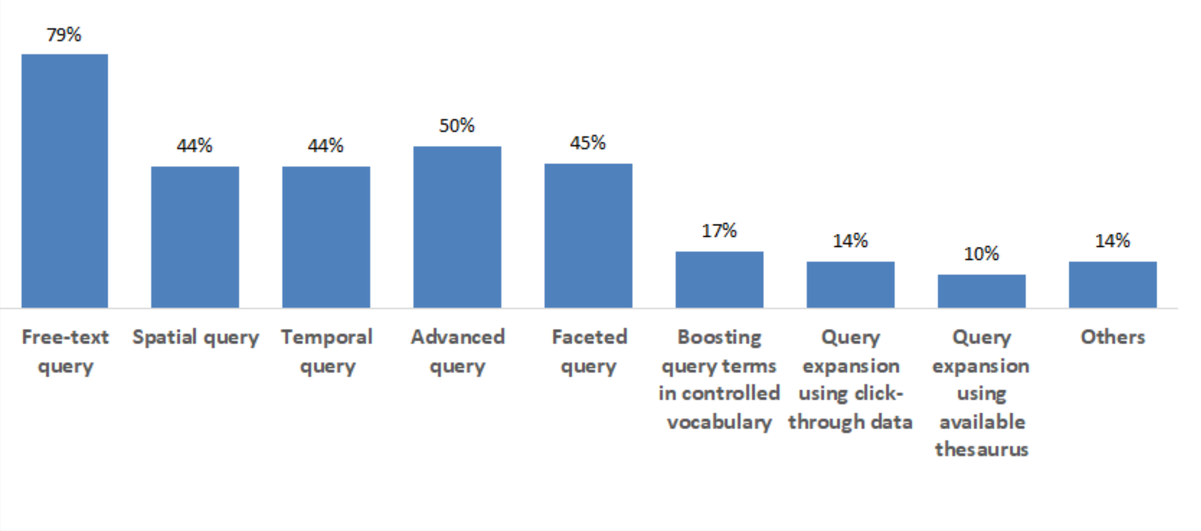

Figure 9. Query construction methods (N = 98)

from earth science and/or physical science domains (67%). Interestingly, out of 9 that support only temporal query, 5 of them cover the social science domain. The “Other” types of queries offered include “standardized parameter names”, “regular expressions”, browse and SQL. Again, most answers from the “Others” are a further explanation of what advance queres include.

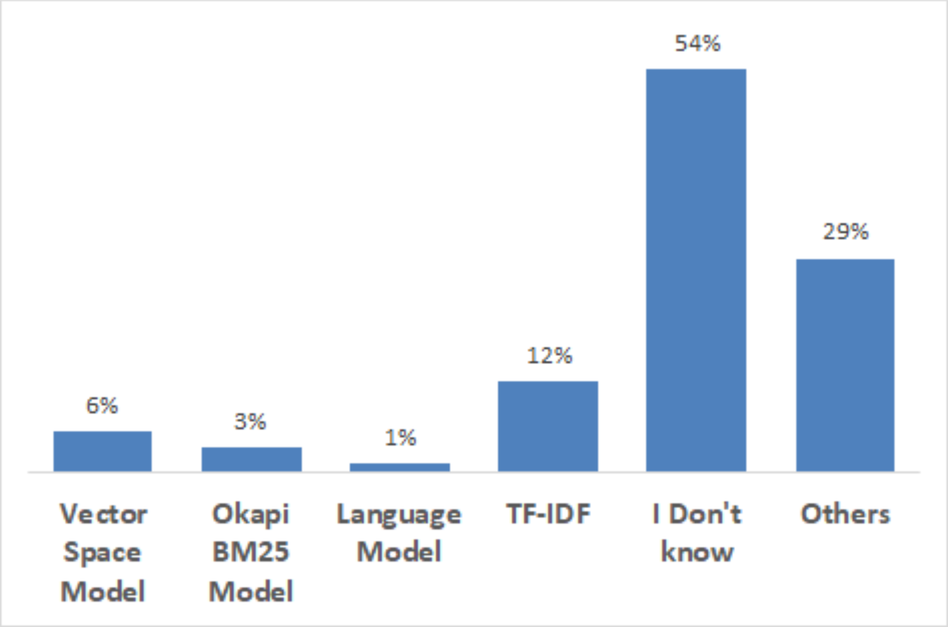

Figure 10. Ranking models (N = 90)

Relevancy ranking model (Q11) Out of 90 respondents who answered Q11: “What relevancy ranking model is used by your search system?”, most responses (54%) are “do not know” (cf. Figure 10). Most likely they just plugged in a chosen system (e.g. Solr or Elasticsearch) and deployed default setting (vector space model with TF-IDF weighting factor), as some answers from the “Others” indicate “We rely on Elasticsearch’s internal ranking model for relevancy” or “Solr standard” without knowing what the default ranking model is. Twelve percent of respondents know their repository deployed the TF-IDF. Note that nearly half out of these answers from the “Others” (13 out of 27) say they do not do ranking or this question does not apply to their repository, while the remainder (14) either rely on their system’s (e.g., DSpace, Elasticsearch) internal ranking or SQL query as their repository is backed up by a type of DB system (e.g., MySQL, MongoDB).

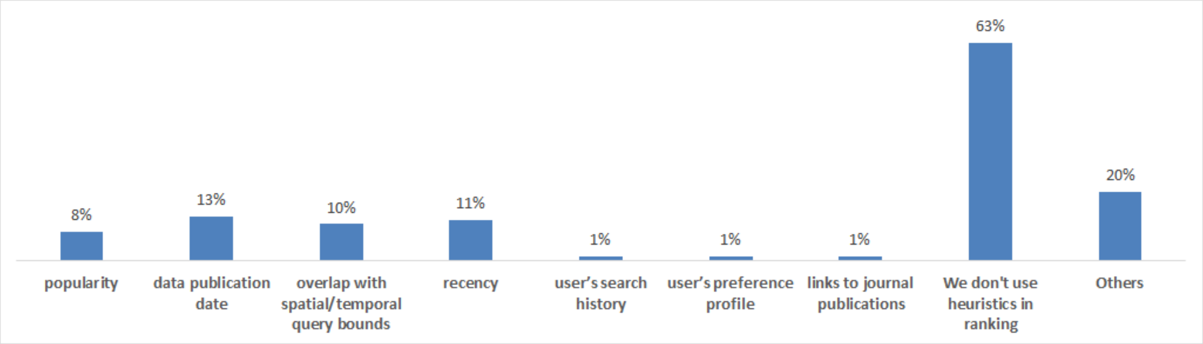

Figure 11. Heuristics incorporated in ranking (N = 96)

Applied heuristics in ranking (Q12) We learned from the test survey that we performed that some repositories used heuristics in ranking, so we included Q12 to ask explicitly about this. Sixty-three percent of repositories did not incorporate any heuristics in ranking (cf. Figure 11). Only a small number incorporated data publication date (13%), recency (11%), overlap with spatial and temporal query bounds (10%) into ranking. The “Other” category included data quality (1), priority ranking (1), most curated/trusted data elements (1).

5.3 Evaluation and Benchmark

Evaluation The main purpose of this survey was to collect information that could help repositories improve their search systems. Thus, it was important to know whether the repositories had conducted any evaluations of their search systems and, if they had, how it had been done and what steps had been performed to achieve improvements. Out of 98 responses to the question “Have you conducted any evaluations to measure the performance of your search system?” (**Q13**), 23% responded Yes, 77% No (cf. Figure 12). Most repositories that had small number of indexed records did not conduct an evaluation: out of 95 responses that provided a valid index size number to the Q6, only 16% of 75 repositories that had less than one million indexed records conducted any evaluation, while nearly 50% of repositories with one million or more indexed records conducted an evaluation.

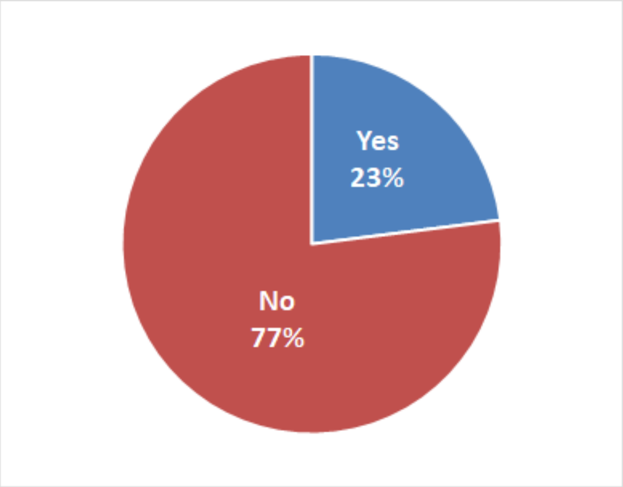

Figure 12. Percentage of repositories evaluated their search system (N=98)

Confidence in search service (Q30) Respondents were asked how confident they are in the quality of their search service: nearly half of respondents (47%) believed that their search engines would satisfy most users (cf. Figure 13), although only 23% of repositories performed some form of evaluation. Other 42% responded they are not sure - they would like to do a systematic evaluation, but 28% said this was not currently a priority for them, 16% said they did not have the necessary resources and/or in-house expertise.

For the 46 who believed their search engines would satisfy most users, 34 of them hadn’t conducted any evaluation yet. 6 out of those 23 who conducted evaluation were not sure about the quality of their search service while 52% of those who conducted evaluations believed their services would satisfy most users.   (**Q13**, **Q30**)

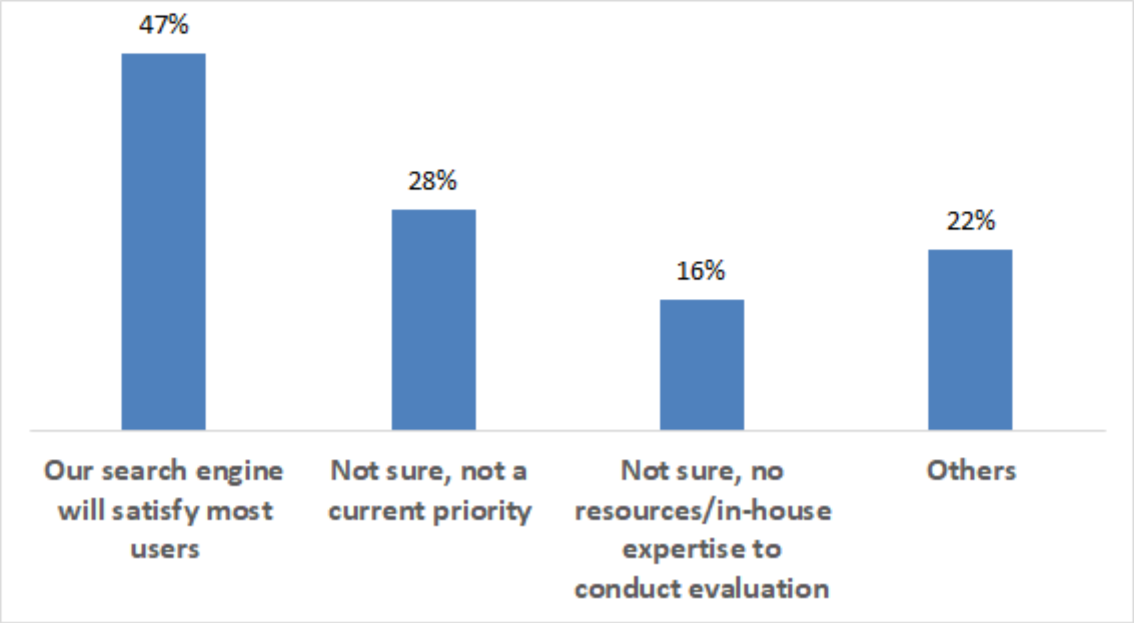

Figure 13. Confidence in search engine performance (N = 98)

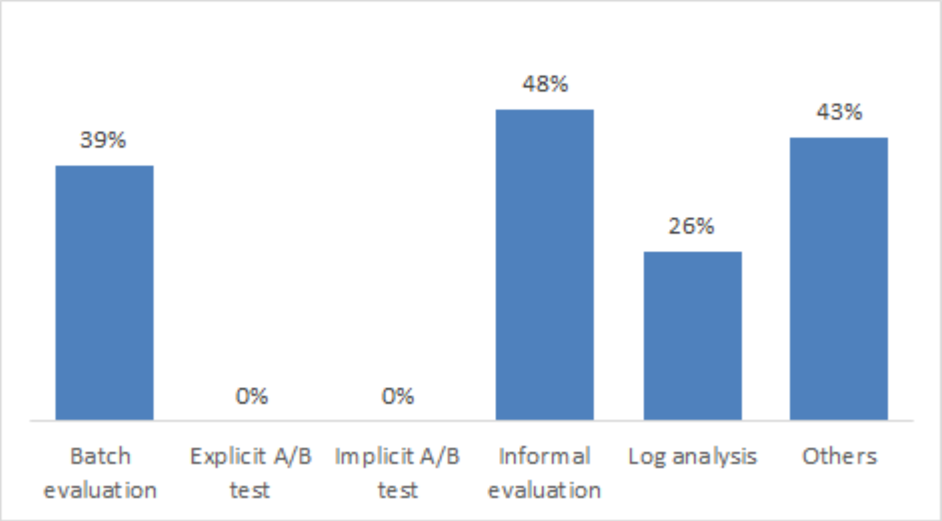

Figure 14. Evaluation method (N = 23)

Evaluation method (Q14) Regarding methods for conducting their evaluations (cf. Figure 14), among those 23 repositories that had conducted an evaluation, 39% of them conducted batch evaluation by creating a test collection and then using this test collection to measure search performance; 48% conducted informal evaluation, i.e., they did not create a test collection, each time picked a set of queries and assessed relevance of a search result per query informally; and 26% did log analysis to assess the rank and relevancy of clicked records. Valuations in the "Other" category (43%) included performance testing (6), user involved evaluation (2), evaluation of common user search scenarios, and subject matter expert targeted evaluation (1). One respondent commented that this question (and all evaluation questions) used terminology that is not understandable to him/her.

It is unclear how better of evaluated systems performed, as none of these repositories provide a performance measure, which was asked in Q21.

Test collection There was a set of questions on the way a test collection was built and size of a test collection. There was a fairly even distribution of how testing queries were derived (**Q15**): by developers (11), analysing search log (7), potential users (7) and data librarians (7). Five repositories recruited potential users to make relevance assessment (**Q16**), while in the “Other” category Subject Matter Experts, software developers and content managers were mentioned. Out of 12 responses who provided an answer to the scale of relevance assessment (**Q17**): 4 applied Graded one (e.g. relevant, partially relevant, and not relevant), 5 binary one (relevant or not relevant), 2 List A is better than List B (note that no one chose the two A/B test methods as listed in Q14, it is likely the descriptions of two A/B tests in Q14 are specific not covering all possible A/B test scenarios), and 1 used “Summary of interview results and observations”. In terms of what resources were drew in turning relevancy ranking (**Q18**): Internal resource, Vendor, and Resource from 3rd party; 11 out of 12 chose internal resource,1 chose both internal resource and vendor.

We asked three questions concerning test collections - number of records in test collection, number of test queries, and number of records that are assessed per query. The size of test collections (**Q19**) varied from 108 to 18M (13 responses). Five out of the 10 responses to the second question (**Q20**) indicated that they had 100 or more testing queries; most of the others had a number of testing queries that was reasonable considering that the relatively small number of test records. The number of records accessed per query (**Q21**) varied widely, ranging from 2 to 100M.

For the 14 who have built a test collection (either complete or partial), three of them indicated willingness to share their test collection and the other 11 did not believe that their test collection is reusable (**Q22**).

5.4 Discoverability from Web Search Engines

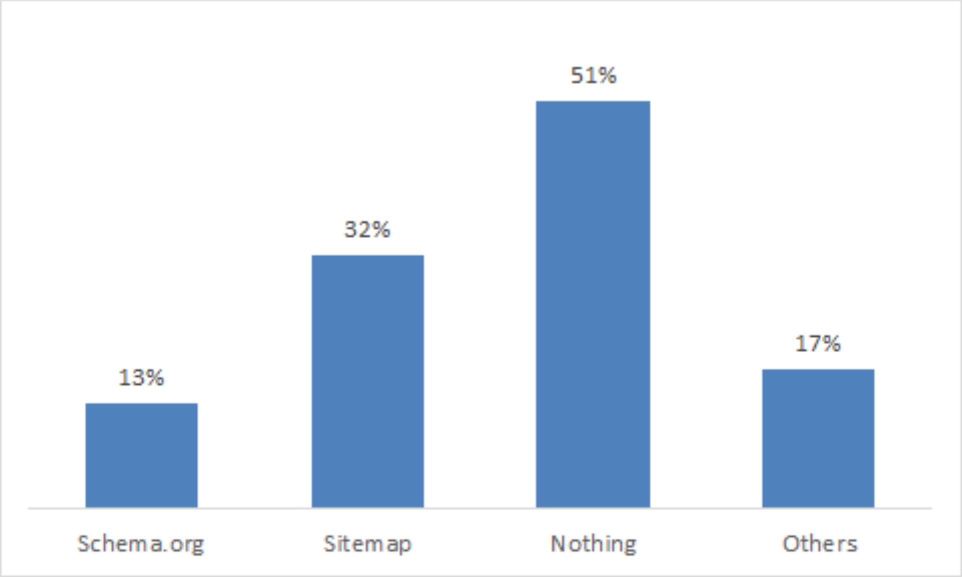

Figure 15. Applied methods to make data discoverable by web search engines (N=97)

About half of the repositories have not done anything to boost their repository records to web search engines such as Google or Bing (**Q23**, Fig 15). More (64%) of those who had performed evaluations of their search system had done something to boost discoverability by search engines. Submitting a sitemap to Google was used more often (32%) than annotating records in Schema.org. Responses from the “Other” included:

- Annotate via metatags;
- Methods as undertaken by vendors;
- Including Dublin core metadata in web pages;
- Implement recommendations for data citation standards;
- Static pages are generated which Google is able to crawl;
- Submit metadata to DataCite;
- Make sure all of our mirror sites point to a single canonical site through a Google webmaster account.

When asked (**Q24**) “Which of your methods have improved your records' ranking in general search engines?” 72% indicated that they had not conducted any evaluation. About 20% and 3% believed that Sitemap and Schema.org had improved their ranking, although it should be noted that some who indicated that Sitemap had improved their ranking had in fact not selected in the previous question sitemap as a method they used. One “Other” response to Q25 was: “Our data catalog is federated by multiple organizations including data.gov, Thompson Reuters Web of Science, etc. We have seen increase in ranking since this started.”

5.5 Other Technologies and Settings

We asked five open-ended questions to elicit what technologies and settings have been adopted to improve index and search.

Index text from publication (**Q25**): Among 11 respondents named technology for parsing text from publications: 9 named Apache Tika (7) and Apache POI (2), they all deployed Solr or Elasticsearch. Others used LibreOffice (PDF parser), Textpresson, in-house developed pipelines, and Microsoft SQL Server CONTAINS.

Index text from Data (**Q26**): Among 15 who named a technology for indexing text from data: five used Apache Tika or POI, five used DB index (MySQL full-text indexing, PostgreSQL, SPARQL), and the rest are largely in-house customisation or implementation.

Machine learning for improving ranking (**Q27**): The question asking whether any machine learning technologies were used for improving relevancy ranking evoked a uniform “None”, N/A or empty. However, one respondent replied to the next question (Q28) that Machine Learning was used to enrich metadata.

Metadata enrichment (Q28): 18 respondents indicated that they used technology to enrich metadata. The responses included (boldface indicates methods used by repositories that did evaluations):

- ➔ JHOVE
- ➔ Machine Learning
- ➔ Extract from data
- ➔ Virtual Observatory’s Unified Content Descriptors
- ➔ Terminology enhancements utilizing SciGraph

- ➔ Perl for marking up text
- ➔ In-house analysis pipelines
- ➔ A Nesstar system is used to markup data which generates rich metadata
- ➔ Custom code (analyse data and augment dataset metadata)
- ➔ Drupal RDF and metatag modules.
- ➔ Feature catalogue
- ➔ from data set itself (geogr. and time extend, contact, ...)
- ➔ In-house enrich metadata via relationships evaluation
- ➔ In-house developed online metadata editor (MEAN stack).
- ➔ Linked Data & Semantic Web
- ➔ Mostly based on manual curation
- ➔ Semantics (ontologies)
- ➔ We do not use such technologies, but we plan to also support rdf-a in our html metadata export handler in the future.
- ➔ "We use enrichment using PubChem, ChemSpider, US EPA Chemistry Dashboard during MassBank record generation with RMassBank (https://bioconductor.org/packages/release/bioc/html/RMassBank.html). The records are available in text files and uploaded to the system. Hence, a script based enrichment of missing meta data or for the imputation of new meta data tags is possible (we use R scripts).

Other technologies and their usage (**Q29**): 12 participants described and commented on other technologies/system settings that they have tried or applied. Responses included:

- ➔ AstroGrid Virtual Observatory search interfaces (Java/Apache Tomcat). Performance was v. poor, did not scale to number of catalogues required.
- ➔ In-house developed metadata editor and search system (MEAN stack for metadata editor, with Elasticsearch for searching).
- ➔ Search uses MySQL database from Fedora 3
- ➔ The atomic line list uses a custom-built search engine from when it was first created. The code has since been rewritten to make it more flexible and accommodate future extensions of the database. So far its performance has been satisfactory.
- ➔ We contribute to the development of BEXIS2. There [sic] Lucene technology was chosen.
- ➔ We have recently completed the 'Mining and Utilizing Dataset Relevancy from Oceanographic Datasets to Improve Data Discovery and Access' (MUDROD) project funded by NASA AIST (NNX15AM85G). This represents the next generation of search which will be deployed at PO.DAAC
- ➔ We link our metadata concerning instruments, parameters, and vehicles to community-adopted, authoritative sources using Linked Data principles.
- ➔ As a first approach we used eXist XML database for metadata and fulltext search. With this huge amount of data it did not work reliably
- ➔ We tried eXistdb. We also had an intern look at Tika to catalogue papers that use our datasets, but we have yet to integrate this information into the site. It was more for a research project
- ➔ Weighting has been applied to specific elements within the DDI XML which improves the relevance of results.

We noted a few possible discrepancies in the responses. For example, in the open-ended response to Q29 the respondent said they applied weighting to improve the relevance of results, but their response to Q13 indicated they had not conducted any evaluation of their system. We also had a few comments about clarity of terminologies used in the survey. Future work could involve a post survey interview to compliment the survey results.

6 Feedback and Suggestions

To identify future activities for this task force that would benefit the intended community, we provided a list of four activities for survey participants to rate (1 = most preferable, 5 = least preferable). They were also able to nominated any other activities (**Q31**). Survey data shows that respondents prefer to 1) Develop detailed recommendations on how to improve relevance ranking using a specific approach and 2) Develop evaluation standards. While “TREC like evaluation method” and “Create test collections and make them available to the community to test their data search system” were least preferred among the four options. This is related to Q30, where 39 out of 94 repositories would like to have an evaluation of their systems, but do not have enough resources or have other competing tasks.

The following is a list of other activities that participants would like the Relevancy Ranking Task Force to work on (**Q32**):

- A main issue here is: what defines relevancy? When is a search result 100% relevant, when is it 99%, when is it 73% etc.? Clarity on the degrees of relevancy and the means to define this is key to this discussion.
- Evaluation of search engine rankings - comparison with peers.
- Facilitate improved relationships with journal publishers
- Intelligent search
- Provide some training materials for relevancy ranking implementation and optimal use of provided solutions by scientists (similar to data and software carpentry training materials).
- Relevancy ranking is not our priority. We would be interested in some data discovery topics, like including primary data into search, using of visualizations to represent results, new concepts of discovery.
- The development of the repository was funded through a federal grant, but the grant did not include nearly the funds required to actually build a repository, nor any maintenance funds. This seems to be a common problem: funding agencies pay only for prototypes or initial bare-bones development of data repositories and ignore the need to fund software development and maintenance, in particular for the many projects that intend to serve a community larger than a single institution or government.
- This questionnaire actually targets other kinds of search solutions than the linguistic corpus search that is the focus of our activities, but it would of course be good if the activities of RDA could be made to include this kind of application, too. We have been discussing how to rank corpus search hits, e.g., in terms of maximally different linguistic contexts for hits (e.g., the "spread" of word usages in the corpora).
- We ingest tokenized lowercase metadata into the Fedora MySQL database. At search time, we get the list of PIDs and hand it over to EULFedora for displaying.

We note that the outputs from other Task Forces created by the Data Discovery Paradigms Interest Group have already addressed some of the above requests. In particular, “Eleven Quick Tips for Finding Research Data” has been accepted for publication in *PLOS Computational Biology*, and “Data Discovery Paradigms: User Requirements and Recommendations for Data Repositories” is being submitted to *Library and Information Science Research*.

Q32 also allowed participants to comment on the survey instrument itself. These were among the responses:

- ➔ Your questions leave out all data collections, which are not mainly textual. In astronomy, or high energy physics, there are many data collections consisting mainly of rows of numbers, and the 'relevancy' you feature in this questionnaire does not really apply.
- ➔ I did not understand some terms in the survey. Many search systems I use do not work too well as they use too specific search terms and rely on bureaucratic descriptions of the datasets and not what scientists use.
- ➔ Some of the technical terms used in survey questions aren't sufficiently defined or introduced (for people whose primary role are of expertise is not search relevancy).

The latter two responses raise two issues: 1) the terminology we used in the survey was not understandable to everyone, which may stem from the fact that the person completing the survey did not design the search system and was not familiar with its internal configuration. Responses to other open-ended questions were in some cases filled with “I don’t understand”, indicating that education on search system basics is needed. Notably, the roles of the people who had difficulty with terminology were mostly manager, curator and maintainer. An alternate explanation is simply that the survey instrument could have provided better explanations of the terminology used. The second response above seems to suggest a misunderstanding regarding the metadata, which is used to describe the data, and the terms researchers use in formulating their search problem.

7 Summary

The following is a summary in response to the five survey question topics:

- Majority of participating repositories deployed either Lucene based search systems or DB based SQL search. (SQT1)
- A system was chosen mainly because it is open source or repositories have relevant skills to deployed the chosen system. (SQT1)
- Majority of systems support free-text query, spatial query, temporal query, advanced query and faceted query. (SQT1)
- Most repositories deploy a system as it is (default setting), do not know which ranking model is deployed or can be modified, and do not apply any heuristics and other technologies to enhance search. (SQT1)
- Less than a quarter of repositories conducted evaluation of search quality, but none of them provide a performance measure. (SQT2)
- Nearly half of respondents believe their search engine can satisfy most users regardless if they have conducted evaluation. Other half who aren’t sure about their search quality, would like to evaluate their system but they do not have in-house

expertise or not treat evaluation as a current priority because of other competing tasks. (SQT2)

- About half of the repositories have tried to boost their repository records to web search engines. The Sitemap is most used method followed by Schema.org. (SQT3)
- A variety of open source and commercial technologies are being used to index text from both publications and data. Despite advances in using machine learning for "learning to rank" (Grotov and de Rijke, 2016), machine learning is not yet being used to improve relevancy ranking, though it is one of several techniques being used to enrich metadata. (SQT4)
- Repositories would like us to develop practice requirements for detailed recommendations on how to improve relevance ranking using a specific approach and evaluation standards, least interested in developing TREC-like test collections. (SQT5)

8 Recommendation and Discussion

In general, people desire guidelines for improving relevancy ranking in their systems. In fact, none of the repositories with >100M records ranked detailed recommendations a 1, and only one ranked it 2. In other words, small repositories are more in need of guidelines. In general, repositories understand that their search systems need to be evaluated and improved, which includes evaluating and experimenting with search models on one's data collections. However, they often lack the resources (time and/or expertise) to explore and evaluate the available options. Not surprisingly, those who indicated a lack adequate resources to do an evaluation (Q30) most often selected "Developing guidelines…" and "Create a community environment for testing search algorithms and providing technical support to each other" as the top priorities. This is perhaps an area for people working in the search space to work on through initiatives similar to TREC and the bioCADDIE 2016 Dataset Retrieval Challenge. It takes considerable effort to build test collections and requires domain expertise. Therefore, a key consideration is to focus on efforts that bring the most improvement at the least cost. It may make the most sense to narrow test collections to specific domains, as bioCADDIE did in the challenge.

This study has examined the practices that data repositories employ in helping users search their holdings and demonstrates the value of looking across disciplines at the common data discovery issues, such as relevancy, that all data repositories face. This, in fact, is the purpose and strength of the Research Data Alliance, which inspired this study.

9 Acknowledgement

This work was developed as part of the Research Data Alliance (RDA) Interest Group entitled 'Data Discovery Paradigms'. We acknowledge the support provided by the RDA community and structures. We thank Minh Duc Nguyen and Christopher Lynnes who provided feedback on an early version of the survey instrument. We also thank members of the Interest Group for their support, especially Jens Klump, Beth Huffer, Dawei Lin, who participated in early video-conference calls on this topic.

10 References

Armbruster, Chris, and Laurent Romary. "Comparing repository types: challenges and barriers for subject-based repositories, research repositories, national repository systems and institutional repositories in serving scholarly communication." (2009).

Baeza-Yates, Ricardo, and Berthier Ribeiro-Neto. *Modern information retrieval*. Vol. 463. New York: ACM press, 1999.

Bankier, Jean Gabriel, and Kenneth Gleason. "Institutional repository software comparison." (2014).

Blaha, Michael, David LaPlant, and Erica Marvak. "Requirements for repository software." *Reverse Engineering, 1998. Proceedings. Fifth Working Conference on*. IEEE, 1998.

Brin, S. and L. Page (1998). The anatomy of a large-scale hypertextual Web search engine. Computer Networks and ISDN Systems 30 (1-7), 107–117.

Emergence, R. D. A. "Guest Editorial Building Global Infrastructure for Data Sharing and Exchange Through the Research Data Alliance." *D-Lib Magazine* 20.1/2 (2014).

Hui Fang, Tao Tao, and Chengxiang Zhai. 2011. Diagnostic Evaluation of Information Retrieval Models. *ACM Trans. Inf. Syst.* 29, 2, Article 7 (April 2011), 42 pages. DOI=http://dx.doi.org/10.1145/1961209.1961210

Fay, Ed. "Repository software comparison: building digital library infrastructure at LSE." *Ariadne* 64 (2010).

Fuhr, N. Goevert, G. Kazai and M. Lalmas. INEX: Initiative for the Evaluation of XML Retrieval, ACM SIGIR Workshop on XML and Information Retrieval, Tampere, Finland, 15 August 2002

Gareth J.F. Jones, Séamus Lawless, Julio Gonzalo, Liadh Kelly, Lorraine Goeuriot, Thomas Mandl, Linda Cappellato, and Nicola Ferro (editors) (2017). Experimental IR Meets Multilinguality, Multimodality, and Interaction. 8th International Conference of the CLEF Association, CLEF 2017, Dublin, Ireland, September 11–14, 2017,

Goker, A., and Davies, J. Information Retrieval: Searching in the 21st Century. John Wiley and Sons, Ltd., ISBN-13: 978-0470027622, November 2009.

Grotov, A, and de Rijke, M. (2016). Online Learning to Rank for Information Retrieval: SIGIR 2016 Tutorial. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval* (SIGIR '16). ACM, New York, NY, USA, 1215-1218. DOI: https://doi.org/10.1145/2911451.2914798

Harman, D. (2005). The Text REtrieval Conferences (TRECs): Providing a Test-Bed for Information Retrieval Systems. In Bulletin of the American Society for Information Science and Technology. Vol 24 (4). 31 Jan. 2005.

Harman, D. (2013). TREC-Style Evaluations. In book: Information Retrieval Meets Information Visualization: PROMISE Winter School 2012, Zinal, Switzerland, January 23-27, 2012, Revised Tutorial Lectures. Editors: Agosti, M., Ferro, N., Forner, P., M{\"u}ller, H. and Santucci, G. DOI: 10.1007/978-3-642-36415-0_7

W., Hersh, A. Turpin, S. Price, D. Kraemer, B. Chan, L. Sacherek, D. Olson, Do batch and user evaluations give the same results? In Proceedings of the 23rd Annual ACM SIGIR Conference on Research and Development in Information Retrieval, pp.17-24, 2000

Khalsa,SiriJodha, Cotroneo, Peter, & Wu,Mingfang. (2018). A survey of current practice of data search services [Data set]. Zenodo. https://zenodo.org/record/1193196

Kökörčený, Michal, and Agáta Bodnárová. "Comparison of digital libraries systems." *International Conference on Data networks, communications, computers (DNCOCO'10), Faro, Portugal. Disponível na www: http://bit. ly/10gajkr*. 2010.

Madalli, Devika P., Sunita Barve, and Saiful Amin. "Digital preservation in open-source digital library software." *The Journal of Academic Librarianship* 38.3 (2012): 161-164.

Mai, J.E., 2016. *Looking for information: A survey of research on information seeking, needs, and behavior*. Emerald Group Publishing. Greengrass, Ed. "Information retrieval: A survey." (2000).

Mitra, Mandar, and B. B. Chaudhuri. "Information retrieval from documents: A survey." *Information retrieval* 2.2-3 (2000): 141-163.

Naouar, F., Hlaoua, L. and Omri, M.N., 2013, September. Relevance feedback in collaborative information retrieval based on validated annotation. In *Proceedings of the International Conference on Reasoning and Optimization in Information Systems ROIS* (Vol. 13).

Norozi, Mohammad Ali. *Information retrieval models and relevancy ranking*. MS thesis. 2008.

Pampel H, Vierkant P. Current Status and Future Plans of re3data.org - Registry of Research Data Repositories. In: Wagner J, Elger K, editors. GeoBerlin2015: Dynamic Earth from Alfred Wegener to today and beyond; Abstracts, Annual Meeting of DGGV and DMG. Berlin, Germany; 2015. p. 287—288. Available from: http://gfzpublic.gfz-potsdam.de/pubman/item/escidoc:1369620.

Pitkow, James E., and Margaret M. Recker. "Using the Web as a survey tool: Results from the second WWW user survey." *Computer Networks and ISDN Systems* 27.6 (1995): 809-822.

Ponte, J. and W. Croft (1998). A language modeling approach to information retrieval. In Proceedings of the 21st ACM Conference on Research and Development in Information Retrieval (SIGIR'98), pp. 275–281.

Robertson, S. and K. Sp ̈arck-Jones (1976). Relevance weighting of search terms. Journal of the American Society for Information Science 27, 129–146.

SMART system: Gerard Salton, editor. The SMART Retrieval System — Experiments in Automatic Document Retrieval. Prentice Hall Inc., Englewood Cliffs, NJ, 1971.

Sparck-Jones, K. The cranfield tests. In K. Sparck-djones (editor), Information Retrieval Experiment, pp. 256-284. Butterworths, 1981.

K. Sparck Jones, ``What's the value of TREC - is there a gap to jump or a chasm to bridge?", *SIGIR Forum*, 40 (1), 2006, 10-20. Pdf

Waclawski, Eugene. "How I use it: Survey monkey." *Occupational Medicine* 62.6 (2012): 477-477.

Ross Wilkinson and Mingfang Wu. *Evaluation Experiments and Experience from the Perspective of Interactive Information Retrieval,* in Third Workshop on Empirical Evaluation of Adaptive Systems, in conjunction with AH2004. pp.221-230. August 23-26, The Netherlands.