Investigate what data search systems and ranking models have been deployed.

Serve as a benchmark to be looked back on in future to assess how much and in what ways data search has improved.

Identify potential collaborative projects from the Survey
1. What are characteristics of each repositories? (5)
2. What are system configurations (e.g., ranking model, index methods, query methods)? (7)
3. What are evaluation methods and benchmark? (10)
4. What methods have been used to boost search-ability to web search engines? (2)
5. What other technologies or system configurations have been employed? (5)
6. Wish list for future activities for the RDA relevance task force (2)
Participants background

N = 98

- Life Science, 35%
- Earth Science, 34%
- Social Science, 28%
- Physical Science, 17%
- Health Science, 12%
- Computer Science, 3%
- All Domains, 16%

- United States (46)
- North America (20)
- Europe (44)
- United Kingdom (7)
- China (6)
- Japan (2)
- Germany (20)
- Australia (17)
- Asia Pacific (14)
- Canada (4)
- Sweden (3)
- Netherlands (3)
- Switzerland (3)
- Italy (2)
- Belgium (2)
- Austria (2)
- Denmark (1)
- France (1)
- Greece (1)
- Hungary (1)
- Ireland (1)
- Norway (1)
- Poland (1)
- Portugal (1)
- South Korea (1)
- Spain (1)
- Turkey (1)
- United Arab Emirates (1)

rd-alliance.org
Survey result highlights ...
Data repositories use common search systems

- Solr, 31%
- Elastic Search, 20%
- Lucene, 18%
- I don't know, 10%
- Google Custom Search, 4%
- MarkLogic, 1%

N = 96

- Others, 29%
- Vector Space Model, 6%
- Okapi BM25 Model, 3%
- Language Model, 1%
- TF-IDF, 12%
- I Don't know, 54%

N = 90
Open source and available skills are top reasons for choosing a search system

- Open source, 52%
- Available expertise & skill, 39%
- Recommended by a trusted source, 10%
- Easier to implement, 18%
- Better support, 20%
- Used by others, 16%
- Others, 35%
- Provided by vendor, 5%

N = 96
Majority didn’t conduct any kind of evaluations

9 Created test collection
11 Informal evaluation
6 Log analysis

No performance measure was provided
Repositories desire guidelines for improving relevancy ranking in their data search system, 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.
The study concludes that there is an opportunity for people working in the search space to collaborate, to build test collections and other efforts that offer the greatest improvements in data search services.

Khalsa, SiriJodha; Cotroneo, Peter; Wu, Mingfang (2018), “A survey of current practices in data search services”, Mendeley Data, v1 http://dx.doi.org/10.17632/7j43z6n22z.1
Thank you

Contact:
fpsom@certh.gr
mingfang.wu@ardc.edu.au
sjsk@nsidc.org

https://www.rd-alliance.org/groups/data-discovery-paradigms-ig