FAIRization: What qualifies a search engine for distributed research data?

Findability of research data has been stated as one of the key demands of the FAIR principles. According to the guidelines, this means

- generating and using Persistent Identifiers for long term identification and preservation,
- generating rich metadata elements for better resource description,
- reusing already well-established metadata standards for better interoperability, and
- registering and indexing of research data to prepare it as a searchable resource, e.g. a web-based search engine.

While these formal requirements may help to set up an infrastructure for a web-based search of distributed research data, they give less guidance with respect to designing an environment for retrieving and discovering this data, which is becoming more and more diverse in terms of provenance, domain, format, type and size.

Concepts, implementations and best practices for search engines for distributed research data are still evolving, the same counts for a corresponding model of Information Retrieval (IR). While IR of literature and documents as primary scientific output has led to a paradigm of searching by means of – with a pinch of salt – keywords, text statistics and facets, this model might not be simply transferred to the indexing and retrieval of research data.

In contrast, our conceptual approach within the DFG funded GeRDI project suggests to support users and researchers in answering their research questions, which require referring to and, if necessary, processing of research data. Consider e.g. a research question like 'What is the correlation between fishery (in different regions) and market prices?' While a text query results in documents matching that query (thus more or less fulfilling common IR expectations), the discovery of underlying or 'hidden' research data may only be accidental and erratic. We therefore suggest making more explicit use of 'proxies' as helpers, transmitters or indicators while searching and discovering research data. A proxy in that sense may be a journal or publication referring to a dataset, a time span, a location, a repository or a statistical office as data provider, a controlled vocabulary, a link to a provenance dataset, or just a colleague or peer – rather than the 'content' of the research data being indexed. While we did not check all of these proxy candidates in detail yet, we are convinced that this kind of contextual information helps to discover research data and judge its relevance in a basically transdisciplinary setting.

In an environment of distributed disciplinary data repositories, most of these proxies are expected to be expressed and encoded as (rich) metadata. Since any of those repositories

operate their own local metadata scheme, it is a basic task to constantly crawl, harvest and transform the metadata to generate and update a unified search index, which supports handling of interdisciplinary research questions. In the poster respectively presentation, we will address some of the challenges and solutions both on the level of metadata processing and search interface, so that retrieval of research data becomes less intuitive and accidental, but more deterministic and transparent.

References:

- Dutch Techcentre for Life Sciences (2017): The FAIR Data Principles explained https://www.dtls.nl/fair-data/fair-principles-explained/
- R. Grunzke et al. (2017): Challenges in Creating a Sustainable Generic Research Data Infrastructure. Softwaretechnik-Trends, 37.
- K. Gregory et al. (2017): Searching Data: A Review of Observational Data Retrieval Practices. arXiv:1707.06937