

Tracking and analysing open research data re-use in scholarly journal papers

In the context of the Open Science movement our research focuses on re-use of digital research data sets provided by research data repositories (RDR).

Researchers are asked to submit their research data as publication supplements to comply with research integrity requirements of funding agencies, research institutions and publishers. The Open Access publication of research data sets furthermore fosters data re-use. Data sharing and re-use has many potential benefits e.g. saving research resources by avoiding the duplication of efforts. Nevertheless the actual amount of data re-used is unknown.

Approaching research data re-use from a Library and Information Science (LIS) perspective we identify several theoretical as well as practical challenges. To give some examples, challenges are starting from identifying a proper definition of re-use according to various scientific disciplines and the relevance of re-use, requirements on the re-usability of research data to the yet under-explored field of data citation possibly indicating the re-use of research data.

Based on a theoretical framework on research data re-use and re-usability we plan to emphasize on data re-use by tracking formal data citation as well as analysing metadata information of referenced data sets.

We aim to better understand (technical) methods and instruments to find and identify research data sets automatically. We differentiate between formal citations (e.g. in a reference lists) and data sets mentioned (informal citation) in the full texts (e.g. methods part, acknowledgements, footnotes) or in data representations (e.g. tables, figures) of scholarly papers. Where formal citations are given, they can be tracked to a limited extent using the Data Citation Index (DCI). Mentions of a data set ID can be found by simple text mining techniques. Some papers only refer to an article that relates to a data set and further describe the re-use in the full text. In these cases the motivation of the citation (e.g. background information, documentary purpose) must be identified. Furthermore we plan to analyse metadata information of referenced research data sets (e.g. via OAI-PMH) available in research data repositories and investigate access, quality indicators and usage conditions, e.g. in order to identify re-used data sets that truly meet the requirements of openness (e.g. Open Access licences).