Data search engines

Transcript of video

By Henning Hansen, UiT The Arctic University of Norway
ORCID: 0000-0002-9186-1441

Date: October 2020

Manuscript:

A data search engine is a tool that will allow you to locate published research data, either entire data repositories or individual datasets. In other words, just like any other search engine, the data search engines will index relevant data archives and direct you to where you will find the requested data.

As a rule of thumb, it is only open metadata that will be harvested by data search engines, and thereby become visible. A search engine will point you to where the data is archived, hopefully with a permanent identifier and link. Depending on which data search engine you are using, the search results generated will differ from search engine to search engine.

Some data search engines specialize in datasets, while others are more inclusive and will direct you to data repositories as well as published articles and research projects. The search engines have the same overarching purposes but have different qualities and functionalities. In this session we will be looking into three of the most commonly used data search engines.

The first data search engine we will have a closer look at is called DataCite. DataCite is a widely used, multidisciplinary search engine that allows you to search for datasets equipped with a DOI as their persistent digital identifier. DataCite gathers metadata associated with existing datasets and whenever a new dataset with a DOI-link is posted, it will automatically become indexed in DataCite. Apart from datasets, DataCite also indexes other kinds of works, more broadly defined, including audiovisual material, articles, as well as data repositories, and also allows the user to search for individual researchers. The search function is, however, far from flawless.

The next data search engine we will look at is Bielefeld Academic Search Engine, commonly referred to simply as BASE. BASE is operated by Bielefeld University Library and provides access to hundreds of millions of documents and datasets. It has similar qualities to DataCite, but harvests metadata directly from thousands of repositories, including DataCite, which means that it will also pick up datasets that are not equipped with a DOI as their persistent identifier.

Just like in DataCite, the BASE search engine makes it possible to search not only for datasets, but for articles, entire books, authors, and research projects, as well as other audiovisual material. BASE offers a much more refined search function than DataCite and is thereby more user friendly, although the search function for BASE is not without flaws either. BASE harvests resources from archives that operate with an Open Archive Initiative. This means that the majority of the indexed documents and datasets are available under Open Access licenses.

The third and last data search engine we will look at is Google dataset search, which was launched in 2018 and moved out of beta in 2020. Google dataset search shares features and qualities with the widely used search engine Google scholar, and it is another service created by the search engine giant and directed towards the global research community, intended to complement Google scholar. The aim of Google dataset search is to unify the many thousands of different repositories for

datasets and make all data searchable, without moving the actual data. Google thereby seeks to address the issue of the fragmented platform for dataset publication.

Google dataset search engine bears striking resemblance to the google search engine we are all accustomed to, but the functionality and purpose is much more specialised. The search engine is a rather user-friendly system, and the fact that it indexes datasets only makes it a little more user friendly than both DataCite and BASE, at least for users who are looking for datasets only. Google dataset search shows where the original data is stored, where it can be accessed, available formats, information about funders, and a short description.

The search engine is compatible with Google scholar, which allows users smooth access to information regarding any citations of the data and any associated publications.