Data archives/repositories

Transcript of video

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

Date: October 2020

Manuscript:

In a separate session we have talked about data search engines, which harvest metadata from thousands of data archives and direct the user to the data archives where they will find the data. We have also talked about the Registry of Research Data Repositories, also known as re3data, which is the foremost indexer of data archives. In this session we will talk about the actual data archives, which are also known as data repositories.

The repository is where the datasets are stored, and you will find the actual data. Data repositories come in all sizes. At one end there are small institutional repositories which store a few dozen datasets, and at the other end the major international repositories, which boast tens of thousands of datasets. There are both multidisciplinary and subject specific repositories and most of them are free to use and specialise in open research data.

Many research institutions have their own data archives, often referred to as institutional data repositories. The same goes for many companies, but the data stored here is often not as easily accessible to members of the public. Researchers are often free to choose if they want to deposit their data in the institutional repository or somewhere else. Most institutional repositories only welcome research data from their own researchers. However, within certain disciplines, there are institutional repositories that welcome research data from scholars worldwide, independent of any institutional affiliation. Once the data has been shared in the data repository anyone is free to access them.

For those who lack access to institutional repositories, such as independent scholars, the independent research repositories constitute a good alternative. Zenodo, Figshare, and Dryad are some of the best-known independent repositories. The independent repositories allow anyone to deposit their data and thereby make it possible for independent scholars to comply with data sharing requirements.

Zenodo is one of the major players among data repositories. It is an open access-repository, which was developed under the OpenAIRE program, and it is operated by the European Organisation for Nuclear Research (CERN). Zenodo welcomes research data within all disciplines from researchers all over the world, and does not have any requirements when it comes to format, size, access restrictions or licenses. The repository provides access to research data as well as articles, conference objects, and other objects associated more generally with research. All submissions to Zenodo are assigned a DOI as a persistent identifier, which facilitates data citation. The citation information is also passed on to DataCite and other data search engines, which makes the data easily findable.

Zenodo collaborates with another major data repository, Dryad. Dryad is another open-access repository which shares a lot of features with Zenodo. In terms of subject specialisation Dryad focuses more on scientific data and medical sciences and is particularly strong when it comes to a number of specific subjects. All data deposited in Dryad are associated with a published article and are available under a Creative Commons Zero license (CC-0), which makes the data public domain. Dryad encourages not only the researchers themselves, but also encourages institutions, journals, learned societies and publishers to share their data.

Figshare is an online open access repository that was launched in 2011. It shares features with both ZENODO and Dryad and operates with creative commons licenses. All datasets are available under a Creative Commons Zero license (CC-0), and Figshare welcomes all kinds of different research outputs, regardless of file format. Figshare is backed by a long line of institutions and research associations, but the repository can be used by researchers regardless of institutional affiliation.

Another example of a major player in the field of data repositories is the Harvard Dataverse, which is open to all kinds of research data from scholars from all over the world, both within and outside the Harvard community. Dataverse is an open source software installation, which is used by research repositories all over the world. All the Dataverses share the same interface and system, but each individual Dataverse installation may have a specific focus.

There are national Dataverses, for example the Dataverse for the Netherlands (DataverseNL) or the Dataverse for Norway (DataverseNO), as well as Dataverses for specific institutions, or even for specific departments within these institutions. Each of these Dataverses may in turn host Dataverses on their own, for specific research groups, and so forth.

All of these data repositories are compatible with the identification system ORCID. ORCID is an alphanumeric code that is unique for every academic author and research contributor. In some cases, researchers are required to register with ORCID in order to submit data to a repository.