EUDAT – European Data Infrastructure

by Juha Hakala

The European Commission has a long tradition of supporting the creation of shared information services. Results have been mixed, but some projects, like Europeana[1], have produced impressive results. The recipe for success is sufficient funding and clear mandate, both during and after the project phase.

In addition, projects must deal with various technical challenges. Development and maintenance of national services such as library union catalogues is challenging enough; extending such systems to European scale may make things difficult indeed.

European Commission and research data

The European Commission has a long tradition of supporting library initiatives. For instance, Europeana was launched a decade ago. In comparison, the commission’s large scale involvement with research data is a recent thing. The first EUDAT (European Data Infrastructure) initiative was launched in October 2011. The second 3-year phase of the project, EUDAT2020, received funding from the European Union’s Horizon 2020 programme and was launched in 2015.

EUDAT is building Collaborative Data Infrastructure (CDI), which “will allow researchers to share data within and between communities and enable them to carry out their research effectively”[2]. CDI consists of a set of services which will be described below. But the project has also established an administrative framework which will support the development and maintenance of these services even after the project phase is over. In June 2016, 17 organizations have already signed the statement of intent to join the EUDAT Collaborative Data Infrastructure Agreement. The members will share the costs of the secretariat, but services will also be funded by customers (such as research institutions), which purchase the services on behalf of the users (researchers) who utilize them.
CDI will be cross-disciplinary; all kinds of research data will qualify. One important part of the project is to establish contact with existing research infrastructures. At the moment, EUDAT is working with seven core communities such as CLARIN and ELIXIR, which are, respectively, “The Common Language Resources and Technology Infrastructure” and “a distributed infrastructure for life-science information”. EUDAT has also established links with many pilot initiatives, including Aalto data repository. From Aalto’s point of view, “EUDAT solutions can play a major role in supporting the implementation of the ambitious research data management goals set by Aalto University”.

Co-operation between EUDAT, core communities and other projects strengthens EUDAT’s mandate as a service provider. Moreover, now that EUDAT has a solid organizational basis which will survive after the project is over, it is likely that there will be more partners in the future. National research data initiatives such as the Finnish Open Science and Research initiative can benefit from co-operation with EUDAT in various ways. For instance, copies of Finnish research data sets could be stored in EUDAT services, and Finnish data sets could be made more visible via EUDAT search service, B2FIND. From a technical point of view the latter is easy since only OAI-PMH harvesting of metadata is required. The latter is more demanding since a common agreement of Submission Information Packages to be sent is needed.

From the mandate point of view, it is interesting that EUDAT is not the only generic research data related initiative the commission intends to support. In April 2016 the commission announced that

*the Commission plans to create a new European Open Science Cloud that will offer Europe’s 1.7 million researchers and 70 million science and technology professionals a virtual environment to store, share and re-use their data across disciplines and borders. This will be underpinned by the European Data Infrastructure, deploying the high-bandwidth networks, large scale storage facilities and super-computer capacity necessary to effectively access and process large datasets stored in the cloud.*

It is not yet clear how the Open Science Cloud will relate to EUDAT and domain specific European (and global) research infrastructures. But CSC and other EUDAT partners are aware of Open Science Cloud and believe that it will use the infrastructure created by EUDAT. Generally, any major system managing research data should be

- open in design, participation and use
- interoperable and
- distributed

Data archives not capable of exchanging data (and metadata) with other data archives would be counterproductive especially in the long term. Cost effectiveness requires the possibility of storing copies of data sets in different archives. Moreover, long-term preservation of research data is only possible with migration, and that would be difficult in a closed system.
EUDAT services

EUDAT’s vision is to enable European researchers and practitioners from any research discipline to preserve, find, access, and process data in a trusted environment, as part of a Collaborative Data Infrastructure (CDI) conceived as a network of collaborating, cooperating centres.

CDI network may contain two kinds of nodes:

- Interoperable nodes must have a data repository in which they preserve or curate data from a single research community, or host data from several communities or experiments. It must be possible to harvest the metadata about research data sets into EUDAT, and the metadata must contain some form of persistent identifier.
- Integrated nodes must meet all the requirements of interoperable nodes. In addition, they must integrate their local data infrastructure with the CDI’s data management services and connect their services to the common CDI service management infrastructure.

The easiest way to establish an integrated node is to use EUDAT service suite:

- B2FIND Union catalogue to support finding research data
  - http://b2find.eudat.eu/
- B2ACCESS Identity and access control management
  - https://b2access.eudat.eu:8443/home/home
- B2SHARE Public data repository for storing and sharing research data
  - https://b2share.eudat.eu/
- B2SAFE Research data replication (for bit level preservation)
- B2DROP General purpose data synchronization and exchange
  - https://b2drop.eudat.eu/
- B2HANDLE Handle prefix registration service and (EPIC) PID services

These services are currently hosted by various EUDAT partners. For instance, B2FIND is developed (mainly) by CSC, but the system is implemented on servers in Deutsches Klimarechentrum.

B2 SERVICE SUITE

http://www.eudat.eu/services
Customers of these services are European researchers, especially those who participate in European collaborative projects. EUDAT services have been built in such a way that they promote open access to research data whenever possible. EUDAT itself does not assert any rights over any of the data it holds, so all data stored within the CDI retain their original rights. Access to data in the CDI is free at the point of use.

In this article it is not possible to present all EUDAT services. I concentrate on B2FIND, which is in some respects comparable to library union catalogues, although there are significant differences as well. Some of the lessons learned in the development of library systems can be applied to research data systems, but there is a reason to be cautious: research data differs from publications in various ways, and this has to be taken into account in system design.

B2FIND

B2FIND database is freely available at http://b2find.eudat.eu/.

According to the B2FIND home page[11], it is a
discovery service based on metadata harvested from research data collections from EUDAT data centres and other repositories. The service offers faceted browsing and it allows in particular to discover data that is stored through the B2SAFE and B2SHARE services. The B2FIND service includes metadata that is harvested from many different community repositories.

As far as I know, B2FIND is the first and only union catalogue of research data sets, combining metadata from many different sources. As such, it is potentially a very valuable resource for research data community and generally for libraries, archives and museums – especially since a decision to add SRU[12] search interface to B2FIND has been made. This means that library systems and other applications which support SRU can be connected to B2FIND via a standard API, which should increase B2FIND usage a lot.

B2FIND uses OAI-PMH for harvesting, but metadata is not Dublin Core or even DCAT, but very diverse since it originates from many different scientific communities which do not share the same metadata format or cataloguing rules. In fact, there is so much diversity that the task of creating a union catalogue for libraries, archives and museums is easy in comparison.

Finna[13], which provides free access to materials from Finnish libraries, archives and museums, only needs to deal with metadata in some well-known standard formats. B2FIND, on the other hand, needs to deal with potentially very large number of communities and projects, which may not follow the same principles in the description of their resources.
The B2FIND user guide[14] says that

The B2FIND repository collects diverse metadata from heterogeneous sources inside EUDAT and presents them in a consistent form. The homogenisation of the community-specific data models and vocabularies enable not only the unique presentation of these datasets as tables of field-value pairs but also the faceted search in the B2FIND metadata portal or via an easy to use command line tool.

The approach chosen by EUDAT has been to make the barrier of supplying metadata to B2FIND as low as reasonably achievable. This makes it easy for the research data communities to get on board, and in my opinion the project does not have other options, especially because there is no standard for describing research data sets. But the flip side of the coin is that homogenization and display of metadata in diverse formats may be difficult. Metadata formats and character sets may have been poorly documented by data producers, or they may have been used in unexpected or wrong ways. Some examples of the way these issues manifest themselves in the current version of B2FIND are given below.

However, we librarians should remember that even our bibliographic databases still contain poor MARC records, after the format has been in production for more than 40 years. We should be patient and give the research data community time to develop solid principles for description of research data sets. There are vibrant communities like Research Data Alliance[15] and DDI Alliance[16] which are developing the standards and other specifications that are needed.

Another challenge B2FIND is facing is that it is an aggregator of aggregators.

B2FIND is harvesting metadata from other aggregators such as Europeana. When the record is modified twice, some of the relevant content may be lost in translation. The option of harvesting metadata directly from the source should be considered whenever possible, especially when it is clear that the first aggregator already simplifies the harvested records.

However, whatever technical issues there are, in my opinion they are secondary compared with the need to decide what is valid from the B2FIND point of view. At the moment, the database contains not only metadata records describing research data, but also metadata about scientific articles, books and even manuscripts. Since users do not expect to find these material types from B2FIND, the benefit of including metadata about material types other than research data in the database may be limited.

Most B2FIND metadata providers concentrate on research data sets. But there are some providers like The European Library or e-lis (e-prints in library & information science), which are not research-data oriented, and it might be a good idea to reconsider their appropriateness as harvesting sources in the future.

I have not made any exhaustive analysis of metadata quality, but some providers seem to encounter some challenges in this respect, especially if the metadata records are old. Since B2FIND cannot improve a bad record, this kind of metadata is a problem, and methods of avoiding harvesting of such records should be investigated.
B2FIND contains also metadata about materials which belong under the research data umbrella, such as code books and questionnaires. Such metadata would be useful, but only if these resources are described in a manner suitable for them, and the resulting metadata is presented accordingly.

Although Unicode has solved many character set related issues, union catalogues which have many data sources may still have problems with indexing and displaying characters correctly. B2FIND is not an exception; there are records in which Chinese characters are not presented correctly.

Although some data providers are challenging, there are also some who seem to be non-problematic. A positive example of successful data harvesting and migration is DataCite; there are 1650 records in B2FIND which originate from DataCite and they are all relevant (deal with data sets) and it seems that the metadata is sufficient and correct. In my opinion, services like DataCite are an indication of things to come in research data. And once all services have reached the same maturity level, it will be a lot easier to maintain B2FIND – and to use it from e.g. library systems via standard interfaces.

B2FIND, like other EUDAT services, is still in the early stages of its development. The project has not had much time or resources yet for data related testing. Technically the system is solid – a CKAN-based search application works well. Feedback from the end users has so far been limited, but it is likely that there will be more of it when the system becomes more popular.

Continuous testing and evaluation, combined with analysis of user feedback, is needed in any publicly available information systems to guarantee that the quality of service is sufficient. EUDAT is no exception in this. The main difference between a library catalogue and B2FIND is that library systems have existed for decades, while research data applications are relatively new, and development of cataloguing rules and formats for research data is still to some extent a work in progress. As long as each project and community may choose to use its own formats and cataloguing rules, creating a good union catalogue may be a difficult task. Therefore it is good news that the DDI Alliance is currently preparing an ISO standardization of its specifications. The alliance is also planning to extend the coverage of its specifications. On the other hand, the Research Data Alliance has already been very useful in e.g. insisting upon the usage of persistent identifiers. These initiatives, and other research data standardization efforts, should be supported also by libraries in order to make sure that metadata about publications and research data sets – and systems containing such metadata – can be interlinked.

The author is an IT specialist at the National Library of Finland.
[1] Europeana portal (http://www.europeana.eu/portal/) contains more than 50 million artworks, books and other resources from across Europe


[10] https://www.dkrz.de/


[14] https://eudat.eu/services/userdoc/b2find-usage
