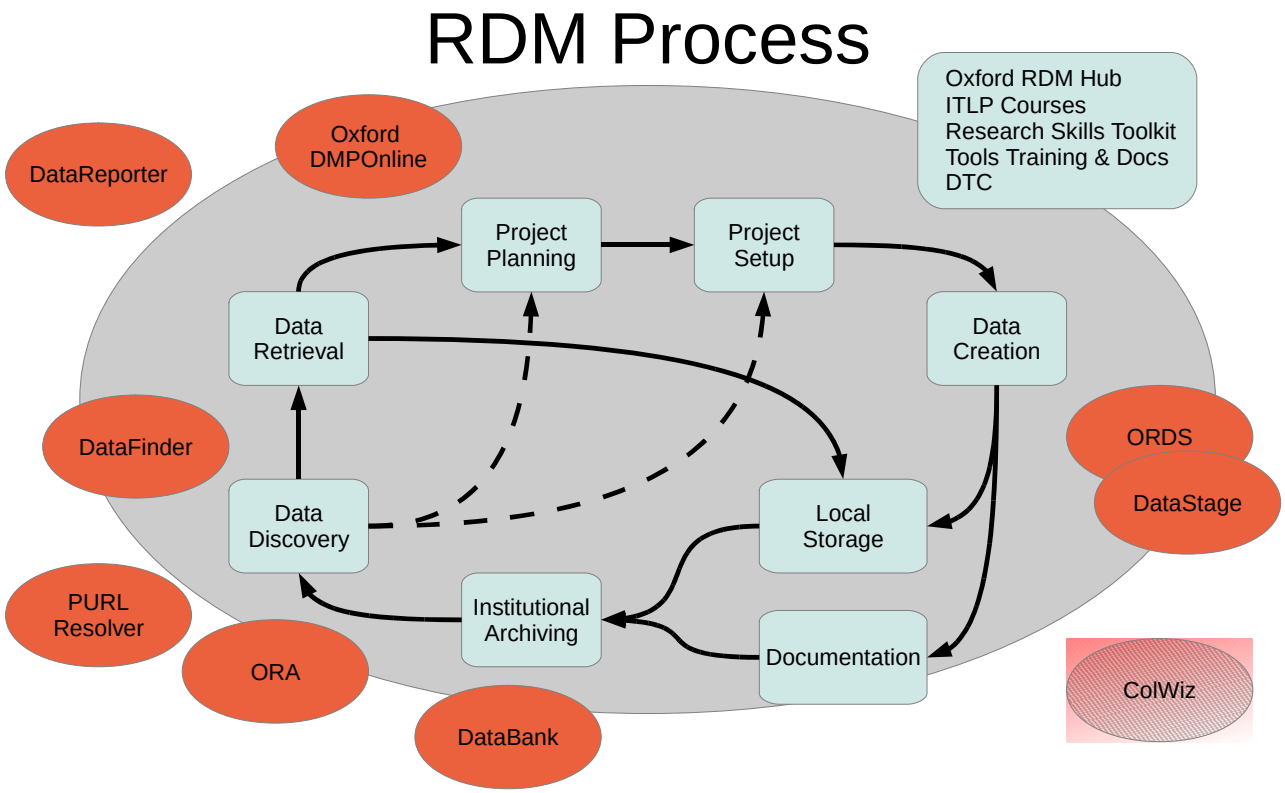
The Great Data Continuum

Just as traditional libraries have had to grow from building their digital catalogues in the 60s and 70s to embracing the web and integrating with other online catalogues, digital libraries are having to move away from being individual information silos to playing the role of data brokers, serving a wider community. In order to serve our users better and to keep up with the federated growth of digital data, we have to integrate our bibliographic data with data from multiple sources within and outside of the university, from individual departments with smaller collections to publishers and funding agencies.

The uptake of any new technology is measured by its ability to satisfy the users' needs, its ease of use, and the value-added services that can be derived from it. To promote usage and understanding of the data being served for consumption, it is necessary to construct a holistic view of the data. There are different facets of information regarding research and research outputs that includes research projects, funding agencies, organisations, policies, and the outcomes generated by the collaboration between the stakeholders involved.

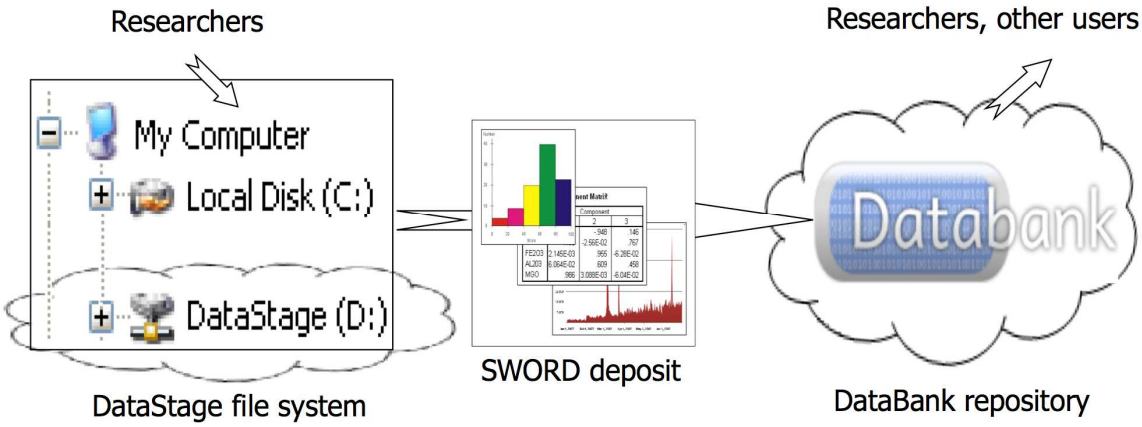
Concentrating on the research community for now, we can broadly classify the data we need to hold and are interested in, into two main areas – publications data and research data. The Bodleian Digital Library has been collecting electronic publications and data about publications for about a decade but have just started the journey into collecting research data and metadata. This paper outlines the flow of information between various systems developed by the Bodleian Libraries in collaboration with other systems owners inside and outside the University. In particular, we will discuss the advantages and challenges of harnessing information from multiple sources and acting as data brokers.



Anusha Ranganathan
Neil Jeffries

UNIVERSITY OF OXFORD

At Oxford, we have adopted a multi-agency tools-based data management approach to enable researchers to plan, discover, work with, annotate, publish, and permanently store their research data. Within the Bodleian, **DataStage** is a secure personalized 'local' file management environment for use at the research group level, appearing as a mapped drive on the end-user's computer. **DataBank** is a scalable data repository designed for institutional deployment. **DataFinder** is a data catalogue that covers both digital and physical information – in a similar manner to our conventional library SOLO service which provides access to physical books and manuscripts as well as electronic journals.



Researchers save files to DataStage just as they would on any other local or shared drive -- but with added extras:

- **Private, shared and collaborative** directories, with password-controlled access
- **Web access** – work with stored files over the web, anywhere in the world
- Users can add **richer metadata** via the web interface
- Users can **invite colleagues** to access group files, via password control
- **Repository submission** interface makes it easy for researchers to define data packages, enter metadata, and deposit them in a repository of choice
- Packaging done using BagIt file packaging specification, and submission is SWORD-2 compliant

When data is ready for publication and archiving, the researcher can then, through a convenient web interface, deposit the selected datasets from their local DataStage system to their institutional or a subject-specific data repository, such as a Databank repository. On submission, Databank will:

- Create a new version of the data-package, with a persistent identifier registered with the University PURL resolver
- Assign a DOI to the data-package and register the DOI with DataCite, enabling the data package to be cited elegantly (provided the submission meets DataCite requirements)
- Serialize the data in the data-package and index it for easy search and retrieval
- Perform archival tasks on the data package

Anusha Ranganathan
Neil Jeffries

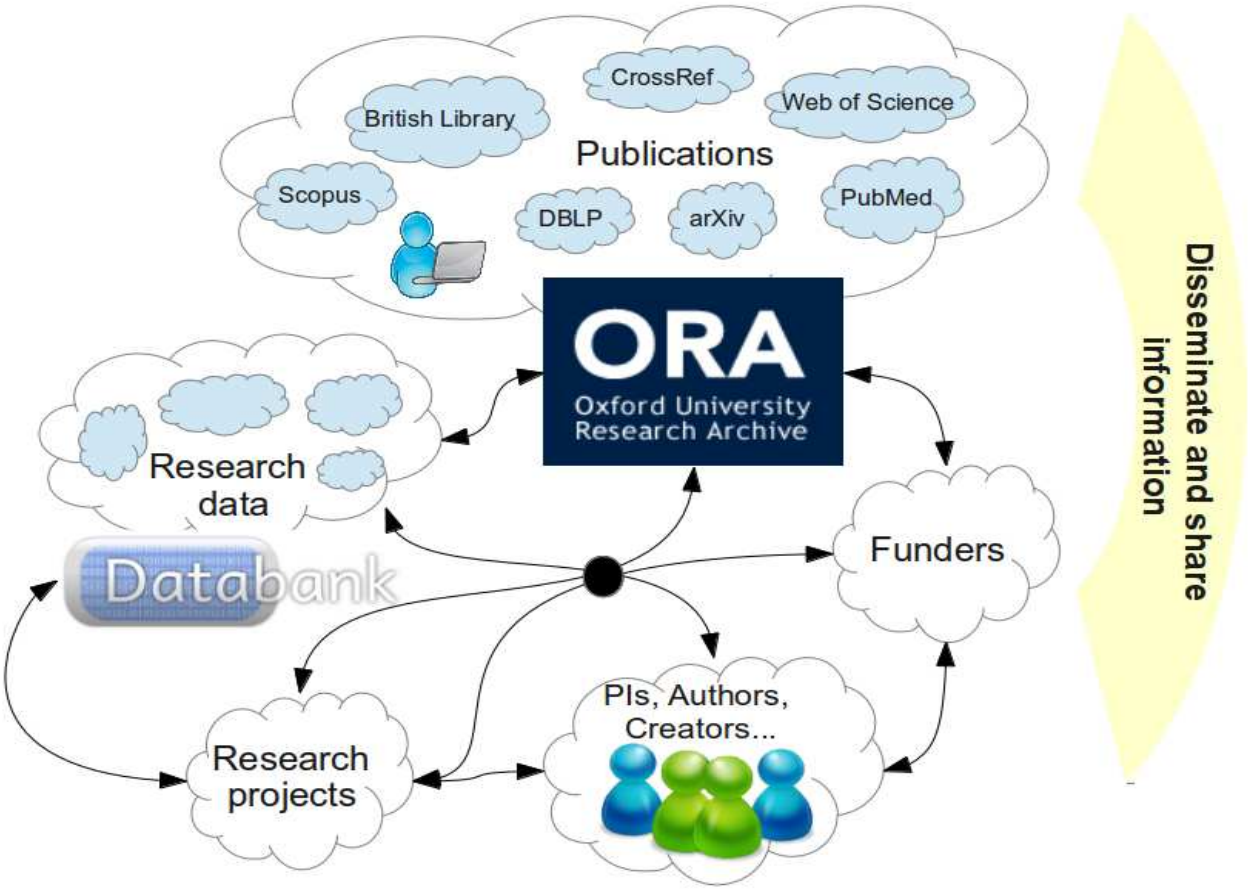
UNIVERSITY OF OXFORD

Researchers can also manage their data in Databank; control access to their data by embargoing it while granting specific users access, publish data as open access, cite their data and link their data to publications held in ORA (Oxford University Research Archive).

Linking research data with publications and sharing the information with other systems

At the Bodleian library, in just over a year we have seen our collection of research publication records grow from a few tens of thousands to well over a hundred thousand items. We periodically aggregate data from multiple repositories within the university as well as from external data sources and share the data with departments within the university. The driver for this has been multi-fold – providing a richer collection of information, better understanding of activity happening within the Oxford research community, and an easier system for enhancing and integrating with research data.

The bibliographic data is enhanced, by adding author links, who are then encouraged to add the publication literature associated to the bibliographic record. This record is further integrated with records containing information on research projects and funding bodies and linked to the research data held in Databank, if available. These resources can now be disseminated to funding bodies, research councils and other departments or services within the university.



DataFinder

Anusha Ranganathan
Neil Jeffries

UNIVERSITY OF OXFORD

Given the heterogeneous nature of the different streams of research and the outputs they produce, it is only natural to accept that not all data can and will be housed and maintained by one system, but several such systems are to co-exist. All of these different sources are fed into Data finder, a finding aid for research data.

We look forward to discussing solutions and workflows adopted by us and other institutions across the world and in gaining insight on some of the challenges we face, especially in maintaining the quality and sanctity of the information gathered.

Anusha Ranganathan
Neil Jeffries

UNIVERSITY OF OXFORD