

The International Journal of Digital Curation

Volume 8, Issue 2 | 2013

Challenges in Building an Institutional Research Data Catalogue

Sally Rumsey,
Digital Research Librarian,
University of Oxford

Neil Jefferies,
Research and Development Project Manager,
University of Oxford

Abstract

The University of Oxford is preparing systems and services to enable members of the university to manage research data produced by its scholars. Much of the work has been carried out under the Jisc-funded Damaro project. This project draws together existing nascent services, adds new systems and services to 'fill the gaps' and provides a wide-ranging infrastructure. Development comprises four parallel strands: endorsement of a university research data management policy; training and guidance in research data management; technical infrastructure; and future sustainability. A key element of the technical infrastructure is DataFinder, a catalogue of Oxford research data outputs. DataFinder's core purposes are to record the existence of Oxford datasets, enable their discovery, and provide details of their location. DataFinder will record metadata about Oxford research data, irrespective of location, discipline or format, and is viewed by the university as a crucial hub for the university's Research Data Management (RDM) infrastructure.



Introduction

The University of Oxford is preparing systems and services to enable members of the university to manage research data produced by its scholars. Much of the work is being carried out under the Jisc-funded Damaro¹ project. This project draws together existing nascent services, adds new systems and services to ‘fill the gaps’ and provides a wide-ranging infrastructure. Development comprises a number of strands: endorsement of a university research data management policy; training and guidance in research data management; technical infrastructure; future sustainability. A key element of the technical infrastructure is DataFinder, a catalogue of Oxford research data outputs. DataFinder’s core purposes are to record the existence of Oxford datasets, enable their discovery, and provide details of their location. DataFinder will record metadata about Oxford research data, irrespective of location, discipline or format, and is viewed by the University as a crucial hub for the University’s Research Data Management (RDM) infrastructure.

DataFinder seeks to provide a consistent and cohesive view of the University’s data assets. However, this view must be constructed from a wide range of sources with data of varying quality and format. In particular, four key cases need to be handled:

1. Data that is stored in machine-harvestable repositories that have sufficient metadata capabilities for DataFinder’s purposes;
2. Data that is stored in machine-harvestable repositories that do not have sufficient metadata capabilities for DataFinder’s purposes;
3. Data stored in non-harvestable repositories;
4. Data that is not online or not digital.

Similarly, the metadata aggregated in DataFinder, along with the usage statistics that it collects internally, are consumed by a number of services besides the basic search capability. The requirements of these services are not necessarily aligned, and include:

- Administrative statistics for evaluation as part of the Research Excellence Framework;
- Evidence for research funders to demonstrate compliance with funding and data management mandates;
- Arbitrary, domain-specific metadata.

This requires a flexible approach to data storage and management as well as creative solutions to indexing, retrieval and formatting. At the same time, the system must be able to cope with the somewhat divergent requirements of significant scalability and longer term sustainability. Thus many design decisions must involve pragmatic trade-offs in order produce a workable outcome.

¹ Damaro: <http://damaro.oucs.ox.ac.uk/>

Metadata

Although a minimum metadata set for research data is being developed to encourage good practice for dataset description at the University of Oxford, designers are forced to acknowledge that it is impossible to impose such a set for all entries in DataFinder. This is because harvested source systems are unlikely to comply with all of Oxford's requirements. It is therefore necessary to balance ideals with reality, resulting in a compromise solution that is fit for purpose, but which operates within significant practical constraints. Such restrictions are the result of human, technical, discipline and policy differences between sources.

The minimum metadata set will be imposed for manual record creation in DataFinder because successful progress through the deposit procedure can be governed by the completion of fields. The set comprises a dozen fields to be completed by the depositor (see Table 1). Additionally, some metadata fields are automatically generated or have default settings. These include metadata rights, depositor information, and a unique identifier for the record. Internally, the Bodleian Libraries uses UUIDs (Unique Universal Identifiers) but the DataFinder system supports the minting and recording of other identifiers that may be applicable to the content, such as DataCite DOIs.

Field to be completed by depositor	Notes
Data creator(s) name (given; family)	Other roles available
Data creator affiliation(s)	—
Data owner	For curation and management
Is the data an output of funded research?	No (no action required) Yes (requires funder name and grant number)
Title	—
Description/abstract	—
Digital/non-digital	Location required (URL or contact details)
Publication year	—
Publisher	Default University of Oxford
Earliest date of access	Embargo set here
Terms and conditions	—
Subject	Initially FAST headings

Table 1. Manual deposit mandatory metadata fields.

The minimum set is based on the DataCite metadata schema² kernel (five fields) and has been augmented to the bare minimum deemed necessary for registration, discovery and citation, these being the key purposes of DataFinder. When compared

² DataCite metadata schema: <http://schema.datacite.org/>

←—————→

against other specialist data services, such as Economic and Social Data Service (ESDS) Data Collection Deposit Form³ or Archaeology Data Services (ADS) Guidelines for Depositors⁴, the set appears light on detail. However, it should be noted that the set has to be generic in order to cater for the wide range of subjects undertaken by Oxford researchers.

Experience with the institutional publications repository has taught us that academics are generally unwilling to complete forms. Over time we aim to capture as much metadata as possible automatically. To begin with, auto-complete will be incorporated when appropriate (for example, creator name, funder). However, it cannot be guaranteed that the depositor is the data creator, or the sole data creator. There is therefore an option to insert depositor details, if appropriate, by indicating if the creator is the depositor, whose details will then be captured automatically.

Additional optional fields allow the depositor to include links to related publications and other online resources that relate to the dataset. They can add a URL, DOI or link to publications held in the institutional publications repository: ORA⁵. Crucially, the system does not proscribe the types of material that may be linked so that related datasets can be accommodated, as well as procedural and methodological documentation that is of growing interest in a number of disciplines.

In the first instance, FAST (Faceted Application of Subject Terminology)⁶ will be employed for indicating controlled subject headings. FAST has been selected as it is broad enough to encompass many disciplines, manageable enough to be used by non-specialists, it is an accepted standard, and there is an open linked data version which will aid interoperability with other semantic resources. Alternative controlled headings, classifications and vocabularies might be incorporated in future to cater for accepted discipline-specific practice. Such vocabularies will need mapping to FAST to enable a consistent baseline for discovery and retrieval. Disciplines with candidate classifications that we are considering include:

- Mathematics (American Mathematical Society subject classification⁷),
- Economics (Journal of Economic Literature classification codes⁸),
- Biosciences (such as those recorded in the BioSharing catalogue of standards⁹).

For interoperability purposes, we are considering a crosswalk to the Joint Academic Coding System¹⁰ (JACS), which is commonly used within academic administrative systems. In addition to controlled classification systems, DataFinder

³ ESDS Data Collection Deposit Form: <http://www.esds.ac.uk/aandp/create/depform.asp>

⁴ ADS Guidelines for Depositors: <http://archaeologydataservice.ac.uk/advice/depositCreate3>

⁵ Oxford University Research Archive (ORA): <http://ora.ox.ac.uk/>

⁶ FAST: <http://www.oclc.org/research/activities/fast.html>

⁷ American Mathematical Society subject classification: <http://www.ams.org/mathscinet/msc/msc2010.html>

⁸ Journal of Economic Literature classification codes: <http://www.aeaweb.org/jel/guide/jel.php>

⁹ BioSharing catalogue of standards: <http://biosharing.org/standards>

¹⁰ JACS: http://www.hesa.ac.uk/index.php?option=com_content&task=view&id=158&Itemid=233

permits depositors to specify free text keywords or phrases that they consider relevant to their content.

In order to promote uptake, the system should adapt to researchers' existing practice wherever possible and not attempt to force them into adopting new ways of working. A number of disciplines have established domain-specific metadata standards which are numerous but overwhelmingly either XML-based or support an XML encoding. Therefore, DataFinder allows researchers to deposit subject-specific metadata by uploading suitable XML files. In practice, it is not practical or scalable for the system to be aware of all the schemas that can be used, and thus effects some consolidation or crosswalk capability. However, indexing tools such as Elastic Search,¹¹ and latterly Apache SOLR¹² are developing schema-less indexing capabilities that mean we may soon be able to provide a meaningful search capability across such domain-specific metadata.

Harvesting and Interoperability

At the beginning of the deposit process, the depositor is asked if they have data to deposit rather than a pre-existing, published dataset. If so, they are then seamlessly directed to a DataBank¹³ deposit form, which appears identical to the DataFinder form but with the facility to upload files. The DataBank record is then made available through DataFinder by harvesting.

The ability to harvest and subsequently manage foreign metadata is a fundamental part of the DataFinder approach. To this end, DataFinder is capable of harvesting metadata using OAI-PMH and provides an OAI-PMH endpoint. As well as permitting interoperability, this approach lends itself to the idea of a tiered DataFinder architecture, where departmentally managed DataFinder instances could feed into an institutional instance and subsequently to a national or supranational level. The use of OAI sets would also allow the creation of specialised catalogues, such as subject or funder specific catalogues.

Looking beyond the repository world, the core metadata in DataFinder will be stored and published as linked open data. Domain-specific XML will not be converted to RDF, but will be provided as a linked resource. In addition to OAI-PMH, a REST API that is closely aligned to the Web User Interface is also provided for machine-to-machine communication. In particular, this is used for DataReporter, discussed later.

Unless the metadata is harvested from an internal Oxford source that the Bodleian Libraries have some control over (such as DataBank), it is highly unlikely that the harvested material will meet the metadata requirements set out in the previous section. Even metadata from subject repositories that have very rich schemas are likely to require some mapping to match terminologies and vocabularies, and in many cases, some manual editing will be required. Therefore, DataFinder has to reconcile potential

¹¹ Elastic Search: <http://www.elasticsearch.org/>

¹² Apache Solr: <http://lucene.apache.org/solr/>

¹³ The University of Oxford's archival research data repository.

updates to a harvested source with a transformed and updated version held internally, and handle the possibility of harvesting more than one record for a given dataset.

The approach taken to resolve these issues is, quite simply, not to attempt an automatic reconciliation. Harvested records are stored as separate objects and updated in step with their sources. These are linked to the augmented record, which acts as the primary source as far as indices and discovery are concerned, and displayed alongside a dataset when viewed (see Figure 1). When a harvested record is updated, the whole set of records are tagged for review so that changes can be propagated in a controlled manner. It is recognised that this approach has the potential to scale poorly but we anticipate that the number of such corrections will remain quite low.

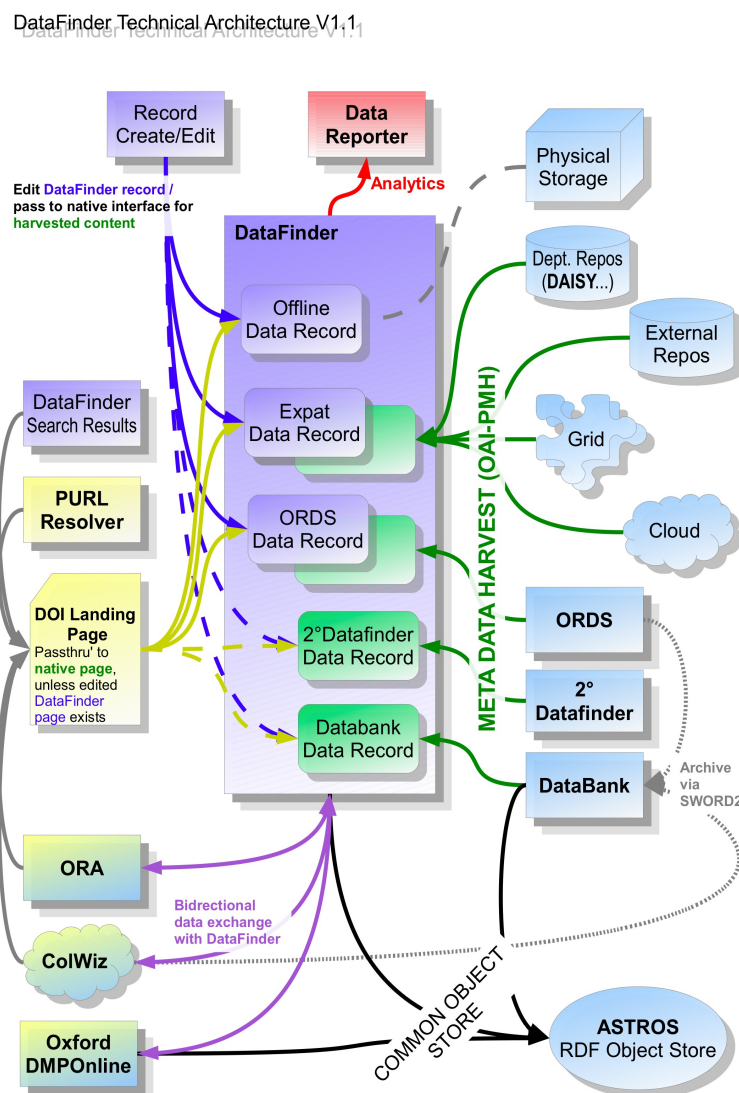


Figure 1. DataFinder architecture.

However, if a source is potentially correctable, DataFinder will attempt to have corrections made at source. Where an augmented record has content that is drawn from a harvested record, it will not permit the editing of that content unless the source has an associated editing URL. Instead, the user will be directed to the relevant URL. It remains to be seen how effective this is.


DataFinder is one of a network of systems and services at the University of Oxford that provide support for researchers and staff throughout the data management lifecycle, as it pertains to the academic process. As the final repository and dissemination point for information about datasets at the end of the process, and the route for discovery of data for new endeavours, it necessarily interoperates closely with a number of services (which are also in development):

- **Oxford DMPOnline:** An online tool for authoring Data Management Plans required as part of funding proposals (based on an earlier DPC tool). These plans can provide DataFinder with some basic metadata in advance, so that the deposit form is partially complete when the time comes to make a deposit. Data Management Plans are deposited in a dedicated section of DataBank with an aligned data model so they can be readily accessed using the same APIs.
- **ORA:** The Oxford publications repository. ORA will support the creation and linking of data records in DataFinder as part of the deposit workflow (and consequent data deposit into DataBank, if required). Similarly, DataFinder will support linking to publications in ORA at the time of data deposit. This will require both resources to exchange metadata regarding their contents.
- **DataBank:**¹⁴ The research data repository service provided by the Bodleian Libraries, which is closely coupled to DataFinder as noted above.
- **Symplectic:** An administrative system that harvests research publication metadata from a variety of external sources. This information is validated by authors and then passed on to ORA. If the author has full texts available then they can be uploaded into ORA. Similar functionality for deposit of related data or data records into DataBank or DataFinder will be considered in future.
- **Oxford Research Database Service (ORDS)**¹⁵: This service provides an institutional private cloud hosted database service for academics with the aim of getting data from local hard drives into a backed up and consistent environment. Databases can be archived automatically into DataBank via the SWORD2 API (and thus have records in DataFinder) at the end of projects.
- **DataStage:** Developed as part of the Jisc DataFlow project, DataStage¹⁶ provides a lightweight data management framework for researchers that, like ORDS, allows data to be archive in DataBank via SWORD2.

¹⁴ DataBank: <https://databank.ora.ox.ac.uk/>

¹⁵ ORDS: <http://ords.ox.ac.uk/>

¹⁶ DataStage: <https://github.com/dataflow/DataStage>

- 
- **Colwiz¹⁷**: A collaborative research environment which consumes information from DataFinder to allow researchers to find and cite data in the environment. However, it can also act as a source of content by allowing researchers to deposit data that is generated within the environment.
 - **Data Reporter**: A reporting system for drawing on DataFinder. Reports can be generated that reflect both content (based on DataFinder metadata) and access (based on PIWIK web logs) statistics. Data Reporter will also generate data exports in formats such as CERIF for research information applications.

User Interface


The DataFinder user interface comprises four main pages: home/search, search results, browse, and deposit. The aim when designing the manual deposit form was to make it as simple to use as possible. The content and functionality of the form is governed by the metadata set described above and by the DataFinder collection and deposit policies. A minimum number of fields are mandatory to ensure deposit is not too daunting, however depositors are encouraged to add richer metadata for superior discovery and to enable users to assess the usefulness of each dataset for their specific purposes. One challenge is to present depositors with appropriate fields depending on the context. For example, a researcher funded by one of the Research Councils UK funding councils is required to display the funder's name and grant number. However, not all depositors should be presented with this element, as it is irrelevant to many. To resolve this, the question asking whether the data being recorded is an output of funded research is mandatory. If the response is negative, no action is needed. If the response is positive, the depositor is required to supply the funder and grant number. Similarly, specific funders' requirements can be made mandatory. For example, if the funding agency is EPSRC (Engineering and Physical Sciences Research Council), then the depositor is prompted to provide documentation about the data (when, why and how the data were generate), which is a requirement of EPSRC funding. For simplification, such documentation might be entered as a URL or other online information. However, this is not future-proof, due to the risk of dead links. A solution might be to harvest information from the link supplied for inclusion as another data stream in the data object.

Records are assigned a status dependent on:

1. Their stage in the deposit process, and
2. The decisions of the reviewer (see Table 2).

Some of the statuses require the staff reviewer to contact either the depositor or a senior reviewer. For manual depositors, a record will not transfer from draft status to submitted status without the completion of mandatory fields.

¹⁷ Colwiz: <http://www.colwiz.com/>




Record Status	Description
Draft	Depositor working on record
Submitted	Depositor has submitted record for review
Approved	Reviewed submission approved without modification
Escalated	Reviewed submission to be checked by another member of staff due to issues, such as commercial or legal agreements, or ethics. Note of problem added to admin record
Referred	More/better information needed before submission can be approved. Submission returned to the submitter with a note of the problem and how to rectify it
Rejected	The administrator reviewing the record has decided that there is something fundamentally wrong with it. Reasons for rejection sent to the submitted

Table 2. Status of records.

Ideally, records harvested from an external source would provide sufficiently complete and accurate records that the harvested material could be made available immediately without further intervention. However, as mentioned previously, it is realistic to expect that a number of sources will fall short of our requirements in this area. These will require further manual enhancement before they can be released. Thus, records harvested from external sources will typically be assigned to either ‘Submitted’ or ‘Approved’ status, depending on the source.

As stated above, the Bodleian Libraries will, wherever possible, harvest existing metadata to populate DataFinder, including from the Libraries’ own DataBank service. We aim initially to work with national, discipline-specific data centres, such as providers, harvesting services such as the Science and Technology Facilities Council (STFC)¹⁸ data catalogue, ADS, and ESDS to gather metadata for Oxford generated datasets. However, sources of metadata for data are not as commonplace as those for publications: researchers have not recently been expected to cite their datasets until relatively. In the early stages, we expect that many datasets will require manual metadata creation, particularly for non-digital data. This begs the question: what will motivate academics to generate such metadata for their datasets, when they have been reluctant to do so for their publications in institutional repositories? The drivers will come partly from the expectations of research funders and compliance monitoring, from journals that demand a link to underlying data, and crucially, if researchers appreciate the benefits of recording datasets in DataFinder for themselves. Alternatively, pressure to record datasets might emerge from individual departments and research institutes, but this is as yet unknown. Researchers might understand the benefits of ‘create once, use many times’ for metadata (for example, that DataFinder metadata can be re-purposed for reporting internally or externally), maximizing the visibility of their research outputs or assigning a persistent URL for citation purposes. The focus for research data management at Oxford is initially on datasets that underpin published articles and those required to meet funders’ requirements. We are therefore hopeful that researchers will be more inclined than not to ensure a record of

¹⁸ STFC: <http://www.stfc.ac.uk/>



their published data exists in DataFinder. This is particularly pertinent for non-digital data, where the online presence is non-existent. There is, at present, no intention of which we are aware for any unit within the University to require its researchers to record datasets in DataFinder.

One option for populating DataFinder is to provide a mediated service where a member of library staff or other individual creates a record on behalf of the data creator. However, this service relies on notification of the existence of data. How such a service might be offered will be explored as the university data support infrastructure beds down.

The approach for the design of the deposit form is based on questionnaire construction. This model aims to make deposit less intimidating and more user friendly than labelling each element with its field title. The field titles have, in the main, been replaced by questions in lay language which attempt to be jargon free.

The aesthetic design of the user interface is based on the current design of the institutional repository. This gives a common look and feel and 'brand' to the services managed and maintained by the Bodleian Libraries. The design of the user journeys for the search and access interface is similar to that of the publications repository.

Conclusion

The University of Oxford has focussed its discussions about support for researchers and provision of research data management systems around a depiction of a generic research lifecycle. The systems being developed have been superimposed on the lifecycle so it is clear where they fit within a general workflow. This is, to some extent, an over-simplification, but has helped to identify key systems, systems that are deemed crucial and need supporting in future, and how the systems interact. DataFinder is identified as part of the 'discover, find, locate' quadrant of the lifecycle and has been identified by the university as the foundation for a well-managed research data management infrastructure. Without it, there is no single means to discover what data exist and where they are located. This expectation has placed a certain amount of pressure on the Damaro project team, and members of the project are keen to emphasise that the first release of DataFinder cannot be expected to resolve all questions and problems associated with research data audit for an institution as large, devolved, diverse and research-intensive as Oxford.

The resulting online service is one that is fit for purpose. There are a number of constraints caused by the problems associated with obtaining rich metadata and in provision of a first release of a complex service. An important part of the university's research data management activities will be to encourage data producers to insert comprehensive descriptive metadata by clearly explaining the benefits of so doing. Even with minimal metadata, Oxford datasets can be recorded, found and cited, thereby making DataFinder a key service in the university's research data management infrastructure.