RDAP Review: Edinburgh DataShare

Edinburgh DataShare – Reflections from a Data Repository Manager

by Robin Rice

EDITOR'S SUMMARY

A no-cost data repository service for university researchers, the Edinburgh DataShare has grown from a demonstration project to an integral part of the research data management (RDM) initiative of the University of Edinburgh. The initiative is designed to serve data management planning, data storage, data stewardship and data support, as well as provide a data asset registry and a vault where data will be retained for a period of time. Pilot testing the repository and RDM services has led to changes in the interface and workflow to improve usability. Testing by a clinical health instructor raised questions about anonymized patient data, confidentiality and vulnerability to hacking and led to a decision to cap the embargo on data access at five years. Additional testing for very large datasets and multimedia datasets remains to be done to determine the minimum amount of data fields that will serve a variety of purposes.

KEYWORDS

digital repositories
data curation
research data sets
evaluation
usability

Edinburgh DataShare is a free-at-point-of-use data repository service that allows university researchers to upload, share and license their data resources for online discovery and re-use by others. It was built in DSpace during the DISC-UK DataShare project (2007-2009) as an exemplar for institutional data repositories.

The project's outputs have become sleeper hits in recent years. Two end-of-project presentations on Slideshare, "Open Data and Institutional Repositories" and "Tackling Research Data in a DSpace Repository," have over 9,000 views to date. Meanwhile, Edinburgh DataShare has transitioned from a project demonstrator to a key service in the rollout of the University of Edinburgh's research data management (RDM) initiative. The university's RDM policy, passed in May 2011, encourages researchers to deposit their data in a university or domain repository:

> *"Research data of future historical interest, and all research data that represent records of the University, including data that substantiate research findings, will be offered and assessed for deposit and retention in an appropriate national or international data service or domain repository, or a University repository."* (*www.ed.ac.uk/is/research-data-policy*)

Robin Rice is data librarian, University of Edinburgh. She can be reached at r.rice<at>ed.ac.uk or @sparrowbarley (twitter)

The university's RDM roadmap (http://edin.ac/XnMS9E) covers four main service areas: data management planning, active data storage, data stewardship and data support. The data repository is aligned with two other stewardship services – a data asset registry, to describe and locate university-produced datasets wherever they reside, and a vault where researchers can store so-called golden copies of datasets that cannot be made publicly available but must be retained for a given period.

The academic-led RDM Steering Group is tasked with ensuring RDM services are fit-for-purpose across the university. They have asked various researchers across the three colleges to pilot the existing RDM services, including the data repository. Each pilot user has challenged us, both in terms of quality of service and policy considerations.

Usability

We observed our first pilot users depositing data into the repository. This exercise was invaluable for improving the deposit workflow and the user experience of the depositor. A number of changes were made, particularly to the hints given for each metadata field. The fields themselves (similar to DataCite) seemed to stand up to the test; users found them useful for creating a complete description of their dataset. However, not all fields applied to every dataset, and users were uncomfortable leaving a field blank, even though only five fields were required. To

CONTENTS

< PREVIOUS PAGE

NEXT PAGE >

reduce confusion, we made a number of changes to the submission workflow and instructions. For example, we now include only one flavor of open data license (Open Data Commons attribution). If users do not select open licenses, they are prompted to fill in a rights statement instead.

Open vs. Closed

A lecturer in clinical health was interested in using the repository with Ph.D. students and their National Health Service (NHS) supervisors in the field for anonymized or aggregate patient-related datasets. We created a depositor's user guide with screenshots to make the deposit process unambiguous and predictable. The lecturers wanted to instill a data-sharing ethic in their students by having a collection for datasets related to student theses. This project was in line with what the students were learning in MANTRA – our open, online course on RDM for Ph.D. students and others. However, there has been resistance from NHS clinicians. In addition to fears of breaches of patient confidentiality, they also have concerns about the press getting stories that could harm the reputation of the NHS.

We discussed options such as a private collection accessible only to the staff and students working on each dataset or the use of a permanent embargo. However, we were uncomfortable with the policy implications of allowing a permanent embargo; in fact, we have decided to hard-code a five-year embargo maximum. This way, users of the DataShare service understand that anything deposited will eventually become public. This policy also avoids the problem of having sensitive datasets locked in a system designed to be accessible over the Internet – which could potentially be hacked. We hope that the active data store coupled with the vault service will meet the needs of users who have datasets that cannot be shared openly. We have also resisted a request to install a registration system for those who download datasets. A registration system would conflict with the ethic of an open repository and place an unnecessary burden on end-users. However, in the future, the data asset registry could be used to provide a metadata record with contact information to request permission for access.

Near Future Releases

There are two more major issues to resolve and test, based on the pilot users' requirements: large and voluminous datasets, and multimedia datasets. Researchers often produce large datasets in fields that do not have a long-term domain repository, even within astronomy and genomics, which are often considered to have well-curated data environments. We are currently testing the SWORD protocol to see if it can reduce the deposit burden on our large data users as well as our own staff. The College of Art has multimedia files that call out for rich display and streaming, as do other fields such as those using medical imagery. We need to find the right balance between catering to these needs and keeping the repository simple and generic. In some cases, the repository will be a sustainable solution for the digital objects themselves, with external websites and databases providing access by pointing to data within the repository. ■

CONTENTS

TOP OF ARTICLE

< PREVIOUS PAGE