Micro Data Repositories: Increasing the Value of Research on the Web

Adam Field and Patrick McSweeney, iSolutions, University of Southampton

9th February 2014

Abstract

As funders and the academic community start to recognise the value in preserving and disseminating data, computing services departments are increasingly called upon to provide infrastructure. We discuss considerations in developing a micro repository approach, where an instance of a repository is created for each dataset, exploring data acquisition and interface requirement. Key to this approach is standardising instances of the repository for reduced support overheads. Details of two micro data repositories are provided as case studies. While the two repositories differ significantly in nature, both have been managed on the same infrastructure and have been well recieved by their respective owners resulting in the creation of an institutional solution.

Background

Initiatives such as the EPSRC's Policy Framework on Research Data[1] have prompted significant consideration into the way data outputs of research projects are made available. A number of solutions have been considered but these tend to focus on a one-size-fits-all approach. Examples include the Open Knowledge foundation's Datahub[5] and UK Data Archive's ReCollect[2] software extension for EPrints. Records in these repositories have the same set of standard fields. The advantage of this approach is that the collection and display of data is straightforward and that preservation concerns, impact analysis and other repository benefits can be realised. The downside of this approach is that the data is not showcased in a manner which provides domain specific value to the community. The researchers who create the data gain very little benefit from its publication. In 2009 Jim Gray identified that in the long tail, science has very poor support for discipline specific creation and distribution of research data[3]. There have been some attempts to address this issue[4] but these have not succeeded in gaining traction at an institutional level. This results in the researcher being responsible for the long term showcasing of their data in a way that is applicable to their research community.

Current data archiving practices

Research projects often produce specialist sets of data that need to be showcased on the web, both for the academic community, and to meet funding requirements. In the past the University of Southampton has approached this challenge using bespoke websites (e.g, fig. 1), archived datasets (zip files on a web site or in publications repositories) and purpose built, script-driven interfaces. However these solutions have proved problematic over time. Archived datasets provide no functionality or added value to the research community. Bespoke websites and script-driven interfaces can add value to the data, but provide only a core set of functionality and usually assume that at point of publication the dataset is complete; these systems are rarely sophisticated enough to provide the capability to add, modify or export records. They are also generally undocumented and have no transferable components, making maintenance a bespoke task.

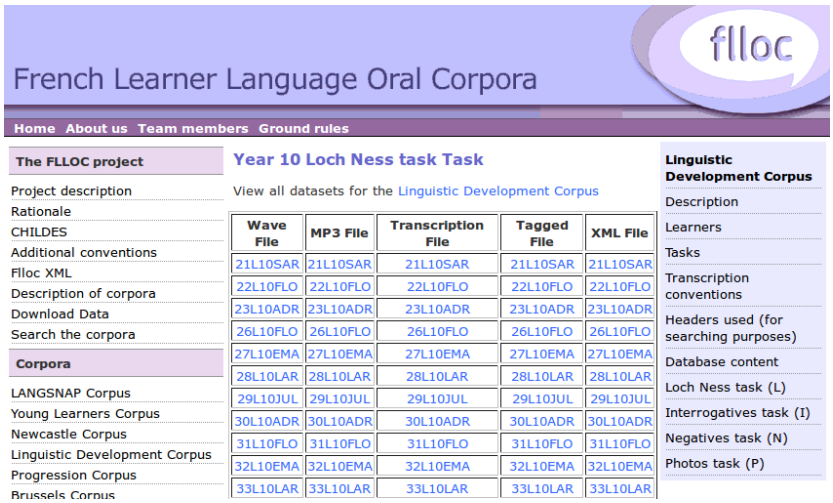

Figure 1: Flloc bespoke website collection view

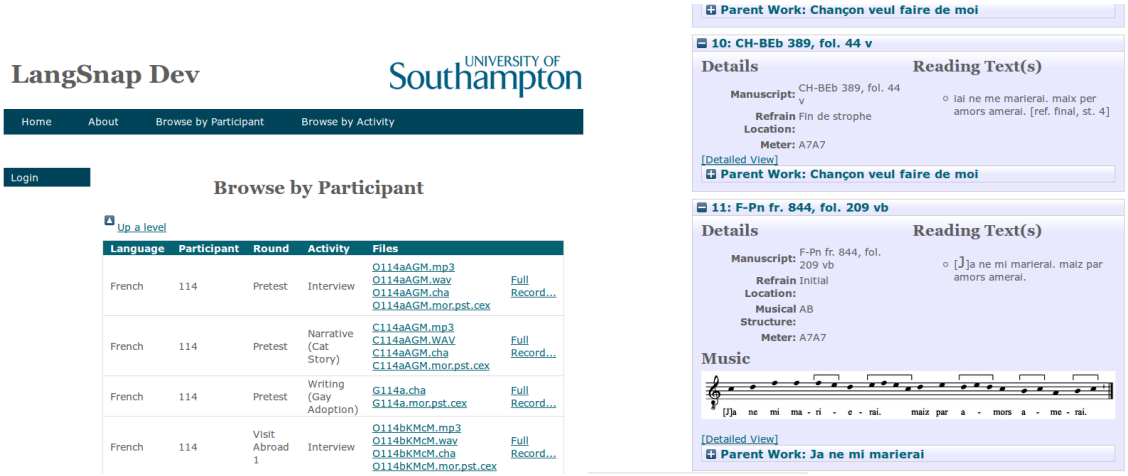

Figure 2: LangSnap and Medmus collection view

Micro data repositories

The increased demand for publishing research data in a way which benefits the community requires a sustainable approach. Using standard repository software to store the outputs of a single project is compelling. The mature code base of a repository platform offers a rich set of tools which would not be written into a one-off bespoke system. Features include browsing, searching, usage statistics and visualising research data. Documentation and popularity of the product helps to ensure that developers will be able to maintain the system in the long term. Furthermore, common export functionality and data standards have the potential to enable interoperability of open research data on the web.

We have produced our first micro repositories with a view to developing a set of best practices to enable a least-effort provision of basic services. This includes creating an on-brand template, developing a consistent approach for managing the configuration and creating standard infrastructure for deploying a repository. Once this infrastructure is in place very little effort is required in creating and managing each new micro repository.

Defining repository objects

The main focus of effort when creating a micro data repository is the preparation of the data. Many researchers do not have an appreciation of data modeling, so existing datasets may not be ready for import into the repository. We have found that encouraging the researcher to think in terms of object types is a good way to start:

- What are the main object types of the repository?
- What fields belong to each object class?
- What are the types, multiplicity and requiredness of the fields?

Drawing up a data model along these lines can form a basis for discussing the specification of the repository as a whole. It is essential that clear data model is achieved before moving forward with the repository build if repeated configuration is to be avoided.

Once an adequate understanding of the shape of the data is reached, the interface and functionality of the repository needs to be considered. A publication is a stand alone object in a publications repository. However, a data record in a data repository tends to be connected to the other records. The way in which the data will be used by the research community needs to be understood in order to determine the best way to view an object in the repository. The same is true for the development of views on collections of data (e.g. a slice of data sharing a particular property), which in a data repository may be far more important than the view on a single data item.

Data repositories are usually constructed around existing datasets, which typically take the form of a spreadsheet, a collection of files or a database. The data can be ingested into the repository either by converting it to a data-standard that the repository supports or using the repository's API to create the records.

Micro data repository case studies

Southampton's micro data repository infrastructure is currently supporting two repositories in the late stages of development (see Table 1). The repositories are quite different in scale and complexity. Langsnap is extremely simple, and follows a fairly traditional repository model of a record. Each item has a handful of metadata fields and a number of audio and/or text files. The Medieval Refrains repository has much larger number of repository objects, an object model that relates the different classes of objects, and some records contain images of staves of music.

The work required to customise the repository interface was a few days in both cases. Many of the practices developed during the creation of LangSnap were successfully applied to Medieval Refrains. Response from the researchers has been overwhelmingly positive.

The future of micro data repositories

Building an infrastructure for provisioning data specific repositories has enabled us to meet the needs of researchers. They are able to engage their community in ways which they were previously unable to and have the capability to curate their data directly. They also get the benefit of compliance with a range of export and import standards, data preservation tools and impact analysis utilities.

Using an off the shelf repository platform has made provisioning new repositories straightforward, maintainable, and requires significantly less staff time than a bespoke solution. This has enabled us to provide more comprehensive support for data outputs using the without requiring extra resources or funding. As a result of this efficiency, we are able to provide more projects with tailored solutions than we previously could have.

The academic enthusiasm for the two pilot repositories has encouraged us to adopt the micro data repository infrastructure as an institutional solution. Provision of a third repository for underwater

recordings of ships is underway and more repositories are expected going forward. Future expansion of the infrastructure includes adding a dashboard for tracking the success of the repositories by aggregating impact information from each one. We are also investigating how outputs can be aggregated with the University publications repository.

Case study	LangSnap	Medieval Refrains
URL	http://langsnap-dev.soton.ac.uk	http://medmus.soton.ac.uk
Data Description	Spoken word recordings and written exercises of language learners before, during and after a year spent in the country of the language of study.	A record of specific spelling of refrains (i.e. fragments of songs) that appear in works (e.g. songs and stories) on manuscripts written in Medieval France.
Previous Solution	Bespoke web page: http://flloc.soton.ac.uk/tasklist.html	Indices in books.
Number of Records	1141	10229
Object Classes	Activity (interview, narrative or written exercise) within a data capture session.	An instance of a particular refrain. An instance of a particular work
Metadata Fields	8	24 and 31 respectively
Data as Provided	Audio and text files with filenames encoding metadata.Audio and text files with filenames encoding metadata.	A word document, which was parsed and converted CSV. The CSV then evolved over a period of 15 months into a final state of two csv files (one for refrains, one for works). Image files of music staves.
Import Process	Script to create repository objects.	Script to create XML from CSV. Import using standard repository tool. Script to attach attach images to repository records.
Micro Repository Complexity	Custom Item Page Custom Collection Pages	Custom Item Page Complex Custom Collection Pages
Customisation Code	github.com/gobfrey/langsnap_eprints	github.com/gobfrey/medmus

Table 1: Overview of implemented micro data repositories

References

[1] Epsrc policy framework on research data. 2011. http://www.epsrc.ac.uk/about/standards /researchdata/Pages/policyframework.aspx.

[2] Essex research data repository and recollect app. 2013. http://data-archive.ac.uk/create-manage/projects/rd-essex?index=1.

[3] Anthony JG Hey, Stewart Tansley, Kristin Michele Tolle, et al. The fourth paradigm: data-intensive scientific discovery. 2009.

[4] Matthew Taylor, Yvonne Howard, and David Millard. Redfeather- resource exhibition and discovery: a lightweight micro-repository for resource sharing. March 2013. http://eprints.soton.ac.uk/349171/.

[5] Mark Wainwright. Opening up scientific data with ckan and the datahub. 2012. http://blog.okfn.org/2012/06/19/ckan-science/.