**The Demonstration Test Catchment Archive: a progress report**

At Open Repositories 2012 we presented an introduction to the Demonstration Test Catchment (DTC) Archive, a data archive intended to facilitate the long-term investigation of water catchment areas in the UK. The current paper reports on the progress that has been made during the last 12 months.

The Demonstration Test Catchment (DTC) programme aims to address the problem of diffuse pollution caused by pollutants which impact on water quality by, for instance, seeping into groundwater and its consequent effects on wildlife: it does this by allowing the detailed monitoring of diverse environmental parameters which together should allow a better understanding of the stressors that produce this pollution.

The DTC Archive, a joint project of the UK Freshwater Biological Association and the Centre for e-Research at King’s College London, is a key part of this strategy. The Archive aims to collect data generated by catchments and make it re-usable in an analysable form by a diverse set of potential users, including scientific researchers, planners, and those who live and work in the areas covered.

The great complexity of this diverse data presents the major challenge to the archive design: it not only includes 'hard' scientific data, such as hydrological information, water chemistry data, biodiversity and ecological data, but also 'soft data' such as surveys, interviews, still images and video. Allowing this diverse set to be accessible and analysable as a single entity is the key requirement of the archive's architecture; this is rendered more difficult by the diverse and unstandardised range of metadata collected for each component. Further challenges arise from the requirement of the archive to conform to the European Union's INSPIRE (Infrastructure for Spatial Information in the European Community) Directive, which mandates the use of specific metadata element sets (such as Gemini2) and markup languages (such as GML (Geography Mark-up Language)).

**Overall architecture**

The overall data model is based on the ISO 19100 series of standards, and comprises a conceptual model for observations (particularly for sampling when making observations) that makes use of the feature types of this standard to enable the ready exchange of observational information units. Two modes of representation are employed to enable the data to fulfil two discrete but inter-related functions, firstly the long-term curation of the data, and secondly its dissemination and re-use.

The first of these representations, which is essentially archival, uses a combination of established metadata standards. The key data files – corresponding to inputs from sensors or from manual observations and samplings – are mostly tabular data such as a CSV files, and each is treated as a digital object with its own metadata. These data objects are described using GML, an XML schema devised by the Open Geospatial Consortium (OGC) to encode geographical features. The objects and their associated metadata are then stored as datastreams within the Archive's Fedora repository, using the open-source Islandora digital asset management system. In addition, taxa are encoded using Darwin Core, a standard for information on biodiversity species information. Other objects in the archive, such as images and PDFs, are described using MODS.

**Controlled vocabularies**

Allowing the re-use of these components as a coherent, analysable data set requires the imposition of conceptual authority control for archive as a whole. Every conceptual component of these data objects (for instance, the headings at the top of a spreadsheet that label the readings recorded in its columns) must be mapped to a central repository of these concepts to identify it unambiguously: in this way, water velocity may be identified as such whether it is labelled 'water velocity', 'water.vel' or just 'vel' in any given resource.

To do this requires mapping each entry to terms in controlled vocabularies, which have been constructed using MADS (Metadata Authority Description Schema), an XML schema from the Library of Congress designed to complement MODS by providing a systematic method for constructing thesauri and vocabularies. The controlled vocabulary is being developed in MultiTes for ease of use while the archive is being constructed. Where possible other existing vocabularies are being used, for example Agrovoc (a thesaurus covering food, nutrition, agriculture, fisheries, forestry, environmental and related domains). Where a term is derived from an existing vocabulary this is referenced.

**Islandora implementation**

The vocabulary has been implemented within Islandora, which is being used to develop the archive. In order to do this effectively a separate tool called the object linker had to be developed. The object linker allows an administrative user to create directional relationship statements for two different digital objects, for instance ‘depicts and ‘is depicted by’, which can be expressed in an appropriate RDF ontology. The object linker then allows a solr query to be written which restricts the list of objects available to either end of this RDF relationship: for the ‘depicts’ relationship, for instance, one solr query would specify that only objects of type ‘image’ can depict things, the other might specify that only people can be depicted. The output of the object linker is an autocomplete pathway that can easily be inserted into a Drupal form so that any field of a metadata form can be linked by means of an RDF statement in the RELS-EXT datastream of a Fedora object to any other object.

This easy ability to link objects makes it far simpler to write a new MADS vocabulary entry in the Islandora system by suggesting terms it may be related to; it also allows the development of a semantically rich set of digital objects in which one digital object may be related to several others according to their relationships in the controlled vocabulary. This approach also lends itself well to publishing these objects or the contents of these objects as open linked data.

**Next steps**

This approach to the data archive allows digital objects to be described at the level of individual projects, datasets and data components, together with the relationships between these various components. While this archival architecture will facilitate data access and preservation at the level of individual files, it is not an effective means of allowing cross-dataset access. To use the data to carry out the intended research, or to reuse it for yet anticipated research, it will be necessary to carry out queries, analyses or visualisations that go across multiple datasets. For example, analysing the effect of farming practices on the levels of a particular substance in water within a catchment will require information originating from a variety of different sensors and experiments.

To support this, the archival dataset representations (which will largely use CSV files) are being mapped onto an RDF-based model, based on the mappings to the vocabulary described above. This is the second representation of the data model referred to above. A benefit of using such a model for data integration is that it is conceptually simple – an “atomic” format that allows users to enter queries and to browse over data from different source datasets and supports cross-dataset querying based on SPARQL. A drawback is that users may find it difficult to write these queries, so we are defining an API that supports various “canned queries”, and supports mappings from this common format into specific forms to suit particular sorts of access, analysis or visualisation, such as, for instance, geospatial representations or time series. An additional benefit is that publishing our data in this form also makes our data part of the broader web of data, allowing other researchers – possibly with other interests – to use it and cite it in their work.