Re-using DSpace to build a repository for freshwater quality data

Andrea Schweer, Stefan Mutter - Information & Technology Services Division
Deniz Özkundakci - Centre for Biodiversity and Ecology Research
The University of Waikato, Hamilton, New Zealand
{schweer,stefanm,denizo}@waikato.ac.nz

This presentation describes how we are adapting DSpace to build LERNZdb, a repository for storing and disseminating New Zealand freshwater quality data. While the original intention in this project was to build a database for individual measurements, the repository model turned out to be a very good starting point for meeting the requirements. This allowed us to re-use our DSpace expertise gained from working on institutional publications repositories. After a quick introduction to LERNZdb's aims, data and users, the main emphasis of this presentation is on our DSpace modifications.

LERNZdb is being developed as part of the Lake Ecosystem Restoration New Zealand (LERNZ) programme at the University of Waikato. The purpose of LERNZdb is to enable LERNZ researchers and others to share, use and re-use New Zealand freshwater quality data. A specific point of difference to other, similar databases is the inclusion of quality assurance information that clearly describes how the data was collected and preprocessed and that consequently enables data users to conduct research in an accountable, reproducible manner.

LERNZdb is designed to accommodate data from two different kinds of sources:

- Manual measurements of water quality are taken in many contexts, for example to support research in the LERNZ programme and as part of ongoing environmental monitoring. The measurement values are typically accompanied by information on the procedures and devices used to collect and create the data. Measurements are usually collected in sets with a common purpose, collection interval and/or geographic region.
- Automated monitoring buoys in several New Zealand lakes measure and transmit data for a range of meteorological and water quality variables. Even though the buoys transmit in near real time, the data received is typically post-processed manually in batches to perform error correction, re-calibration of sensor values etc.

Both types of data can be represented as datasets (that share common characteristics) comprising individual data points.

LERNZdb allows LERNZ researchers, staff members at regional government bodies and other individuals to submit datasets. Following a review process to ensure quality and consistency and an optional embargo period, datasets are made public. Each dataset in LERNZdb is assigned a unique persistent identifier to facilitate citation of the dataset in academic publications and cross-links to any associated publications are shown. End-users of LERNZdb can find datasets using a range of criteria. In a second stage, LERNZdb will allow users to create custom datasets by recombining selected data points from multiple datasets.

DSpace already provides a good fit for the dataset-level requirements but needs to be extended significantly to accommodate the requirements at the level of individual data points. The following describes our completed and planned DSpace customisations for LERNZdb.

Submission process

The LERNZdb submission process strikes a balance between two factors. On one hand, submission needs to be as fast and uncomplicated as possible, to encourage widespread uptake of LERNZdb. On the other hand, all measurements need to be accompanied with sufficient quality assurance and provenance information to allow meaningful re-use and scientific scrutiny. The DSpace submission process allows for customisation through cus-

tom steps. We kept the default steps of descriptive metadata, file upload including embargo as well as license assignment but inserted additional custom steps:

- The submitter describes their dataset by choosing values from predefined vocabularies: What type of data? What (type-dependent) attributes are present? This also defines quality assurance information desirable for the data points in the dataset.
- LERNZdb generates a spreadsheet with columns for measurement values, variable, unit, quality assurance information and other identifying information such as the water body identifier, geo coordinates of measurement site and timestamp.
- The submitter downloads the spreadsheet, pastes in their data and uploads the completed spreadsheet to LERNZdb, along with any raw data files they would like to store in LERNZdb.
- LERNZdb checks the completed spreadsheet for presence of data and compliance with the column format. All information for which we have a controlled vocabulary is cross-checked against the vocabulary; if any discrepancies are found, the submitter is informed and given a chance to revise their spreadsheet. We are planning to extend these checks to include computation of a quality score based on the presence of quality assurance data, with a minimum quality level needed for the submission to go through.

Data model and storage back-end

We use the DSpace community/collection structure to organise datasets according to the supplying organisation. Each dataset record is stored as a DSpace item; the descriptive metadata provided during submission is supplemented with some metadata extracted from the completed spreadsheet. The completed spreadsheet and any additional raw data files are stored as DSpace bitstreams in appropriately-named bundles. On ingest of the completed data spreadsheet, LERNZdb also generates a metadata file in a custom XML format which is stored as an additional bitstream. The XML file allows us to represent structured hierarchical metadata that would not fit into DSpace's key-value metadata model.

In the second stage of LERNZdb development, we will develop an ingestion step that extracts the individual data points from the dataset and stores them in a custom storage back-end.

User and machine interfaces for dissemination

The standard DSpace XMLUI user interface with Discovery enabled already supports search and faceted browse at the dataset level, especially once we configured Discovery to use the metadata fields most meaningful in the LERNZdb context. We may at a later stage add spatial and temporal visualisation and navigation. Dataset record pages – DSpace item pages – are loosely based on those in the Data Dryad[1] and include specific instructions for citing the dataset and any associated publications known to LERNZdb.

OAI-PMH can be used to harvest dataset records, perhaps with a custom metadata prefix to expose the information from our custom XML files. We are also investigating crosswalking dataset-level and datapoint-level information into geospatial dissemination formats.

In the second stage of LERNZdb development, we will create a user interface for datapoint-level operations, either as an extension of the DSpace user interface or as a standalone companion service.

[1] http://datadryad.org/