Session 7: Linked Repositories – Theme of the Day
Moderator: Barbara Butler

AEDA, a Unique Data and Information Management System For the Environmental Sciences

Dr Michael R. Haft
(ORCID ID orcid.org/0000-0002-9584-7882)
Freshwater Biological Association,
The Ferry Landing, Far Sawrey
Ambleside, Cumbria, LA220LP, United Kingdom.

Abstract

The Agricultural and Environmental Data Archive (AEDA) combines the advances in Open Linked Data with Research Data Management approaches to manage a variety of digital objects from documents, images and video to GIS layers and scientific datasets. The subject focus of the Freshwater Biological Association (FBA) initiative is inland waters, their catchments and the agricultural and other environmental influences on their biology, chemistry and ecology. AEDA consists of a data model that meets the needs of long-term digital curation whilst complying with the requirements of the EU INSPIRE Directive on data sharing and compatibility. AEDA also uses a specific controlled vocabulary in order to ensure that all data and other digital information stored within it uses a common language and can therefore be published as Open Linked Data and made available via AEDA's Linked Data API (currently in development). AEDA represents a combined data and information archival and publication platform.

Keywords: Digital curation, data publication, open data, linked data, freshwater biology, environmental science.

Introduction

The Freshwater Biological Association (FBA) was founded in 1929 as a government funded research station. Now a registered UK charity, the FBA has in the course of its existence amassed a large volume of research material on the subject of freshwater science in general, though with an obvious focus on the biological. The FBA library contains approximately 350,000 catalogued items, many of which are exceedingly rare or unique. The FBA is also home to approximately

1200 archive boxes of physical research materials ranging from datasets, to correspondence, to scientific samples. Of particular note is the FBA's Fritsch Collection of algal illustrations. Collected over more than 100 years, the Fritsch Collection contains referenced illustrations of millions of freshwater algae, a unique resource for biological identification.

Figure 1. The FBA Library.

Figure 1. From the FBA Archive, Water Samples from 1920s Expedition to Lake Titicaca.

In 1999 the FBA embarked on a digital project called FreshwaterLife; this was an attempt to lay the foundations of a digital future for the FBA's data and information services. The digital world was fast approaching in the form of websites, online journal publications and data sharing and management. The FreshwaterLife project laid the foundations for future work carried out by the FBA in the digital arena; this led to subsequent funding by the UK's Joint Information Systems Committee (JISC) for the Freshwater Information Sharing Network project (FISHNet) with our partners at King's College London (KCL) and the Freshwater Linked Data project (FISH.Link) with our partners at KCL and the University of Manchester (UM). The results of these projects in turn led to funding from the UK Department for the Environment Food and Rural Affairs (Defra) funding the Demonstration Test Catchment Archive project in order to take the existing knowledge and software developed and build a fully functional digital data curation and data publication system to support the need to preserve and make government funded research data publically available for the foreseeable future. This project was completed in early 2015 and the finished system was called the Agricultural and Environmental Data Archive (AEDA), http://www.environmentdata.org.

AEDA

The Agricultural and Environmental Data Archive uses a unique data model for archiving and publishing digital datasets. This data model is derived from the ISO 19100 series of standards, in particular the Observations and Measurements Standard (ISO 19156:2011); our archive data model can thus be described as a unique profile of this ISO standard.

The data model consists of three primary classes: Activity, Dataset, and Data Component. An activity represents project related contextual information such as responsible party, project start date, end date, contact details and so on. The Dataset class represents a collection of metadata fields to describe a collection of Data Components and other resources. This collection is deliberately left to the discretion of the individual user so as to avoid having to strictly define the nature of a dataset and force users to comply with a definition they may not find suitable for their own data. The Data Components can be thought of as the "payload" of the Dataset. The primary components can be a Measurement, Analysis, Synthesis, Simulation, or Literature Review, all of which consist of numerical data in comma separated value (CSV) format. In addition to the primary data components the model also allows for supplementary files such as PDFs, Word documents, images, Excel spreadsheets etc.

The Controlled Vocabularies

AEDA also consists of a series of three controlled vocabularies used to properly describe the content it contains. They are: the subject thesaurus (www.environmentdata.org/vocabulary), the geographic authorities list (www.environmentdata.org/geographicterms), and the taxonomic authorities list (http://www.environmentdata.org/plist/taxon). These three controlled vocabularies are maintained by the FBA. The vocabularies are used to keyword metadata entries for items in the archive but their most important use is as the column headings in the numerical data files submitted in AEDA Data Component CSV files.

AEDA CSV File Format

The CSV files associated with the primary data components are required to be in a particular format for use in the archive, if the files do not conform to the correct format they are rejected during the file upload process. The format requires that the first three rows of the CSV file be as follows:

- Row 1: Labels, free form and at the discretion of the user.
- Row 2: Observed Properties, must be drawn from the controlled vocabulary and should reflect the most specific parameter being measured by the data, e.g. water temperature should be used in preference to temperature if that is what is being measured.
- Row 3: Units, the unit in which the parameter in row 2 was measured; if no unit is appropriate (as in pH for example) then N/A should be entered. Preferred units for given terms are recorded in the entry for that term in the vocabulary.

Adherence to the above format allows for the issues of ambiguous data recording to be solved; i.e., does t mean time or temperature? It also allows the data to be described in a consistent manner across all Datasets and Data Components, meaning that data can effectively be published as Linked Open Data (LOD).

The Publication Process

When users wish to submit their data for publication they do so from the page associated with the Activity class for their data; publication takes place from this point so as to allow it to cascade down the tree from Activity to Dataset to Data Components. The initial stage in the publication process is for the user to submit data for a “pre-publication check.” This automated check tests their data against the data model and ensures that all the relevant fields have been completed in order to meet the minimum metadata requirements for the data. If the data pass the pre-publication check, they may be passed to the Repository Manager for scrutiny.

The Repository Manager’s job in the data publication process is to ensure that the quality of the metadata and other aspects of the data are of a sufficient standard for publication; if they are not, the data are sent back to the user with recommendations for improvement. Examples of the sort of things the Repository Manager will be looking for are: non-meaningful titles, e.g. RT1, RT2 instead of River Temperature Measurement 1 etc.; insufficiently descriptive abstracts or data quality statements; errors in dates and data formats; or incorrect uses of the controlled vocabulary to describe the data being measured, e.g. using the term Mass to describe Sediment mass. Some of these practices will pass the automated check but are nonetheless not appropriate usage for the data.

If the data are passed as being appropriate by the repository manager and can be published, all digital objects in the scope of the activity involved are cloned and copied to the public area of the repository, where they can then be viewed by the general public and their content downloaded. A Digital Object Identifier (DOI) can then be minted from Datacite (www.datacite.org/) for each dataset.

Any subsequent changes to the data in the system can then be re-published using the same procedure and a new clone of the digital objects is made and assigned a different DOI. These different versions share the same landing and display page but it is possible to navigate between versions on this page.

Because the data have been ordered in this way, it is possible to query them semantically using a query to the archive's triple store. This opens up the possibility of recombining data in novel ways for use in a variety of applications.

Future Work

AEDA will continue to be supported and developed by the FBA for the foreseeable future. We are already working with several scientific research projects to store their data and make it accessible, and we also have an ongoing commitment to make items from the FBA's own physical collections available digitally via the archive.

In addition to the above the FBA has just begun a formal collaboration with the Smithsonian Institute to contribute code and development expertise to their Sidora digital archive project.