2017

Beyond Discovery: Cross-Platform Application of Ecological Metadata Language in Support of Quality Assurance and Control

Jon Wheeler
University of New Mexico - Main Campus, jwheel01@unm.edu

Mark Servilla

Kristin Vanderbilt

It was clear at the outset that the project could not provide direct access to the data for data protection reasons. Via the web portal, CLOSER Discovery (http://discovery.closer.ac.uk), users are able to see what data is available in an accessible way. They can assess the utility of the data for their research and see the full context in which that data was collected. They can extract lists to enhance data access and navigation.

Once this phase of the project is complete, the flexible nature of DDI-L and the software platform will allow us to simply add more information: for instance, the coding used in constructing derived variables, and the provenance of questions to further enrich the contextual information available.

Beyond Discovery: Cross-Platform Application of Ecological Metadata Language in Support of Quality Assurance and Control

*Jon Wheeler, Mark Servilla, and Kristin Vanderbilt**

To support research data curation, descriptive and other types of metadata schemas may be broadly applied to administer access and reuse policies, define system requirements, or perform quality assurance and control functions. In this context, domain repositories like the Long Term Ecological Research (LTER, https://www.lternet.edu/) Network's Provenance Aware Synthesis Tracking Architecture (PASTA, https://github.com/lter/PASTA) are designed to capitalize on complex metadata schema such as the Ecological Metadata Language (EML) to perform an array of descriptive, technical, provenance, and other repository functions.[9] However, transferring data between these and more domain-agnostic systems, such as university institutional repositories (IR), can result in a loss of features when complex metadata are mapped to a more general-purpose schema, such as Dublin Core (http://dublincore.org/). For example, whereas

data sets within PASTA are indexed for faceted discovery across topics including taxonomy, methods, or habitats, mapping related EML fields to Dublin Core results in a conflation of these and other attributes into a single "subject" field. Through spring and summer 2015, a collaboration between the University of New Mexico (UNM, http://library.unm.edu/) Libraries, the Sevilleta LTER program (http://sev.lternet.edu/), and the LTER Network Office (LNO, http://lternet.edu/sites/lno/) explored methods for archiving data sets with complex metadata into an IR. This brief case study describes the application of a standards-based metadata ingest process to facilitate data description and transfer across systems. By establishing and preserving documentation of EML conformance as a baseline requirement for data file properties and metadata syntax, the outcomes to date demonstrate the application of EML as a quality assurance and control resource across the data life cycle.

In 2003, an LTER Network–wide effort to better preserve and expose its data to domain scientists and the broader ecological research community led to the adoption of EML as the network's official metadata standard and the LTER Network Information System (NIS), central to which is the LNO-developed PASTA data repository.[10] PASTA is based on a service-oriented architecture design pattern[11] and exposes an open web-service end-point for data producer and consumer applications, including the LTER Network Data Portal user interface (https://portal.lternet.edu). All data packages submitted to PASTA must be described by an EML science metadata document. Because EML is an expressive metadata standard, the architects of PASTA were able to capitalize on the content and data models defined by the schema in order to enforce consistent data management practices across the LTER Network.

With the EML requirement in place, publication of Sevilleta field data in PASTA is a mediated process providing for information manager oversight and review of submitted metadata against established best practices. Scientists at the Sevilleta LTER submit their metadata to the Sevilleta information manager via a Microsoft Word template, in which they describe who was involved in creating the data set and metadata, where and how the data was collected, what each variable represents, and structural details of the data file. The metadata are then entered into the Sevilleta's instance of the Drupal Ecological Information Management System (https://www.drupal.org/project/deims). DEIMS is a web-based system for managing information products associated with an LTER site.[12] It is implemented in the Drupal content management system. Significantly, DEIMS includes a web-based metadata editor that translates the complexity of the schema into a series of user-friendly forms. Each form represents a subset of the complete metadata, such as sites, methods, people, variables, and data file structure (see appendix 5.0 B). The data, in CSV format, are also uploaded to DEIMS. Moreover, DEIMS includes a custom module that generates EML metadata that is compliant with PASTA's quality control process.

Data files and metadata are logically combined into a "data package" and uploaded to PASTA either manually by the Sevilleta information manager or automatically through DEIMS. As part of the "upload" process, PASTA analyzes the data package for compliance with LTER data management best practices by performing a series of quality checks that compare the descriptive components of the EML to the physical data. A compliance report is then generated by PASTA and is available to producers and consumers of the data package. Data packages that do not comply with critical best practices are rejected by PASTA. Compliance validation of the data package includes asserting the presence of temporal and geographic information, scientific methodologies, designation of field and record delimiters, uniqueness of data attribute identifiers, connectivity of data URLs, and, in the case of tabular data, validity of declared data types and cardinality of the table. Incorrectly recording a string variable as an integer, for instance, will cause the "data type" check to fail. In this case, the Sevilleta information manager would have to correct any errors before the data package can be successfully uploaded into PASTA. Both the EML metadata and data are stored directly in PASTA to ensure direct accessibility for consumers. For data package discovery, PASTA uses Apache Solr to index metadata attributes like key words, creator names, and temporal and geographic information, and provisions a DOI that is recorded by DataCite (https://www.datacite.org/). PASTA also takes advantage of the EML syntax to enable linked open data within the system so that users may embed linked provenance metadata to other data packages in PASTA that were used as source material during synthesis or the creation of derived data products.

In coordination with the Sevilleta LTER and the LNO, the UNM Libraries are providing an archival mirror of Sevilleta data (https://repository.unm.edu/handle/1928/29608), originally published in PASTA, within the University's DSpace-based (http://www.dspace.org/) IR, LoboVault (https://repository.unm.edu/). Archived data sets are harvested from PASTA and packaged per the Simple Archive Format specifications published by DSpace (https://wiki.duraspace.org/display/DSDOC5x/Importing+and+Exporting+Items+via+Simple+Archive+Format) for batch ingest into LoboVault. Using a desktop workflow to coordinate harvest, packaging, and upload into DSpace, the content and metadata included in each package are structured to emulate selected LTER Network Data Portal features using the data package's EML metadata. Specifically, geographic coordinates are mapped to a Darwin Core (http://rs.tdwg.org/dwc/) extension of the DSpace metadata registry and used to draw item-level maps, and a preferred citation is generated that includes the DOI of the harvested data package. While not directly published as item record metadata in LoboVault, the provenance metadata and ingest report described above are likewise harvested for inclusion as downloadable content files associated with their respective data packages. Finally, though the intellectual content of the

EML metadata exceeds the scope of the item record metadata in LoboVault, on harvest a data package's EML is serialized into HTML and likewise included as a downloadable content file. Additionally, because HTML text is fully extracted and indexed within DSpace, the content of the full EML record is thereby exposed to search and discovery features. By using EML metadata in combination with other data package components harvested via PASTA, the archival LoboVault collection supports the long-term curation of Sevilleta LTER data and carries forward the documentation of quality control processes performed by the Sevilleta information managers and within the PASTA architecture.

Summary of Step 5.0: Descriptive Metadata

5.1 Create and Apply Descriptive Metadata: Structure author-generated metadata into the metadata schema used by your repository in order to maximize search and discovery functionality. Create and apply new metadata for the data record, including technical and provenance metadata.

5.2 Consider Metadata Standards for Disciplinary Data: When appropriate, structure and present metadata in multiple schemas to facilitate discovery and future integration into other systems.

Appendix 5.0 B: Screenshots from the Sevilleta LTER Program's Instance of the Drupal Ecological Information Management System (DEIMS)

Jon Wheeler, Mark Servilla, and Kristin Vanderbilt

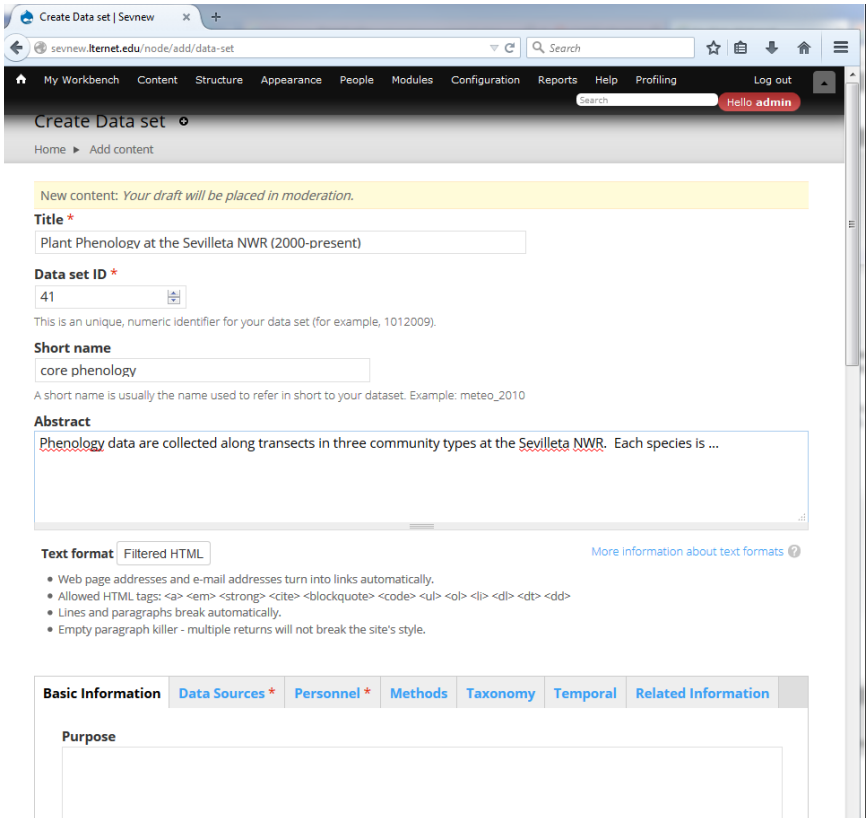

FIGURE 5.5
The DEIMS form to enter discovery level information about the data set captures the title, abstract, and data set identification number. The tabs at the bottom of the screen are used to enter more detailed information, such as methods, temporal and spatial domain of the data set, personnel associated with the data set, and keywords.

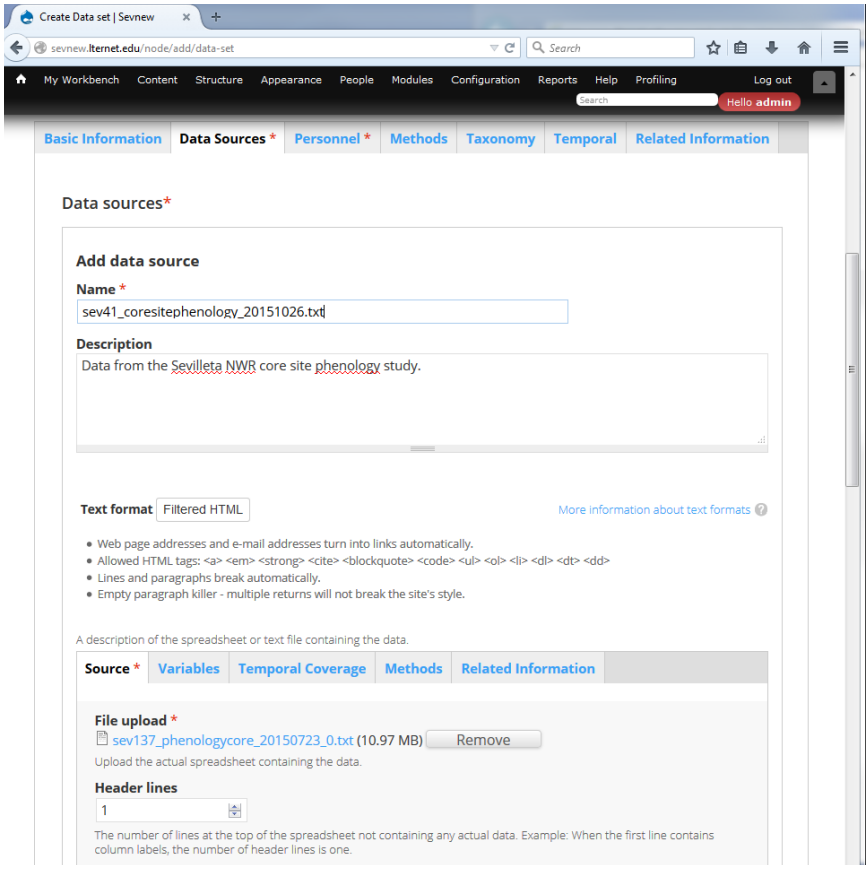

FIGURE 5.6
Data files are added to the data set via the Data Sources tab. Here the user can specify structural aspects of the data, such as the number of header and footer lines, the total number of data lines, and the field and line delimiters.

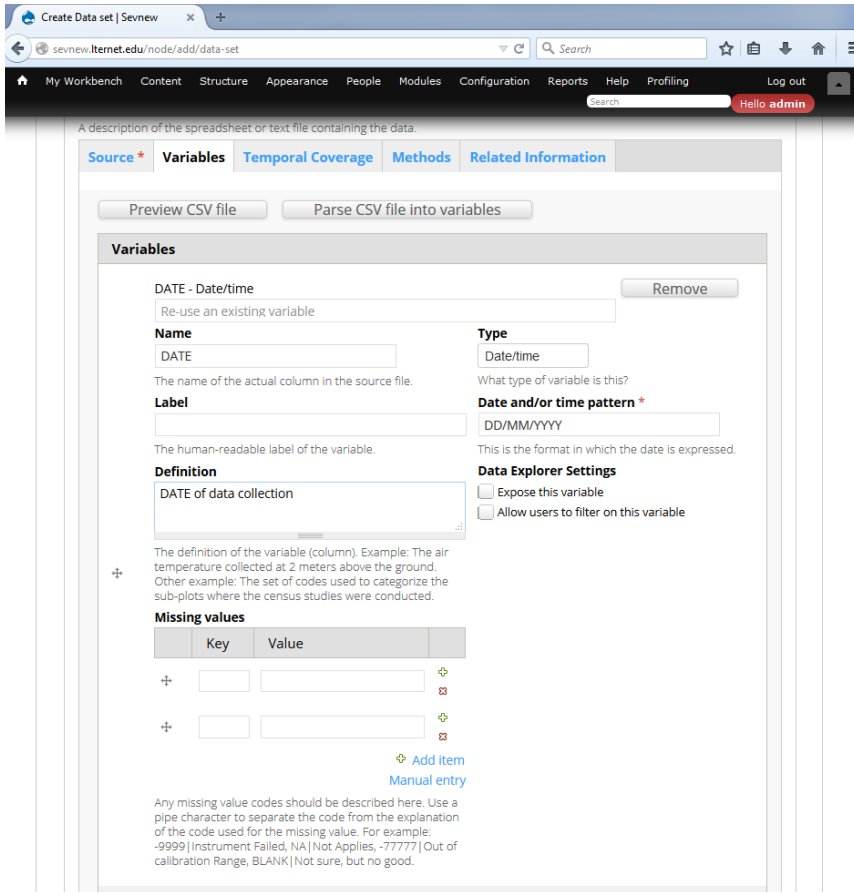

FIGURE 5.7
On the Variables tab, the user can click the button 'Parse CSV file into variables," and the variable names are autopopulated in the form. The user then enters the type of variable (date, nominal, ratio), specifies special formatting (as for dates), and enters the definition of each variable. The code used for missing values can also be entered.

FIGURE 5.8
The new dataset record is entered into the the DEIMS data catalog for the research site. EML can be automatically generated for the dataset.

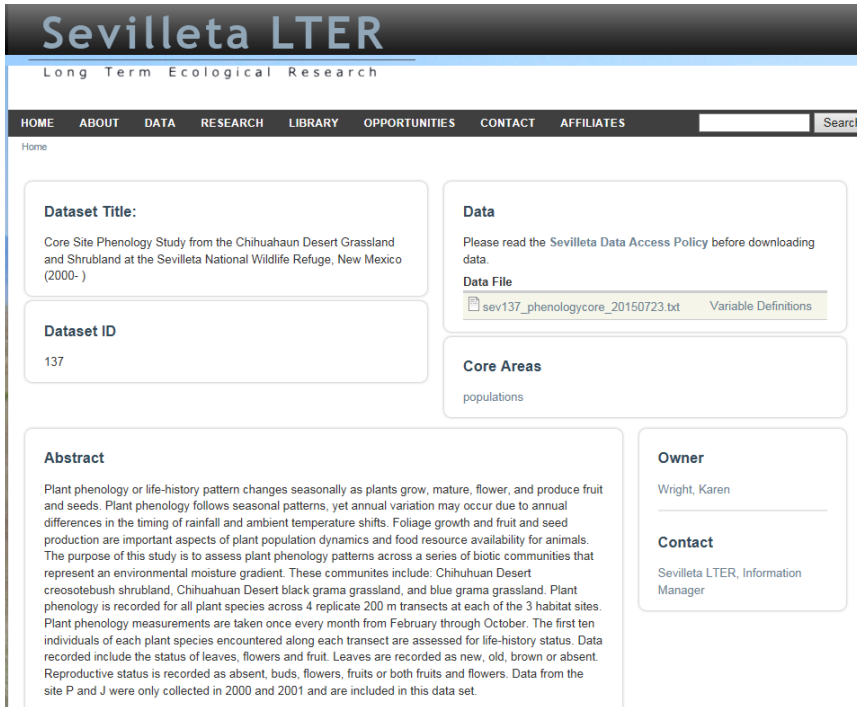

FIGURE 5.9
Built-in features of Drupal can be used to manipulate the display of the metadata and link to the data.

Notes

1. Digital Curation Centre, "General Research Data," accessed March 12, 2016, http://www.dcc.ac.uk/resources/subject-areas/general-research-data.
2. Dublin Core Metadata Initiative. "Dublin Core Metadata Element Set," version 1.1, accessed March 15, 2016, http://dublincore.org/documents/dces.
3. Mary Kurtz, "Dublin Core, DSpace, and a Brief Analysis of Three University Repositories," *Information Technology and Libraries* 29, no. 1 (March 2010): 40–46, doi:10.6017/ital.v29i1.3157.
4. The full metadata profile for DRUM is published online as University of Minnesota Libraries, *The Supporting Documentation for Implementing the Data Repository for the University of Minnesota (DRUM): A Business Model, Functional Requirements, and Metadata Schema* (University of Minnesota Libraries, 2015), http://hdl.handle.net/11299/171761.
5. The methodology and results of the usability testing were previously published as Lisa R. Johnston, Eric Larson, and Erik Moore, "Usability Testing of DRUM: What Aca-