2014

Environmental Data Store: Design And Implementation

Peng Ji

Michael Piasecki

ENVIRONMENTAL DATA STORE: DESIGN AND IMPLEMENTATION

PENG JI (1), MICHAEL PIASECKI (2)
(1): Environmental CrossRoads Initiative, City College of New York, 160 Convent Ave., New York, NY, 10031, USA.
(2): Dept. of Civil Engineering, City College of New York, 160 Convent Ave., New York, NY, 10031, USA.

Abstract: In this paper we present the design and implementation of the Environmental Data Store (EDS). We also highlight the Environmental Thesaurus Server (EnvThs), a controlled vocabulary service application developed by us, providing semantic support on submission and search within EDS. With the rapid growth in data volumes, data diversity and data demands from multi-disciplinary research effort, data management has become a challenge as has the access to diverse but related data for example the context of a project. The EDS is based on the idea to store the 6 data types researchers are likely to encounter in 6 dedicated nodes, coupled via common metadata catalogue and a single web-based access interface, which is built on the DRUPAL platform. Using open source software only the EDS provides repository services for the six fundamental data types: a) Time Series Data, b) GeoSpatial data, c) Digital "Any" Data, d) Ex-Situ Sampling data, e) Modeling Data, and f) Raster Data. Discovery of data in the EDS is supported by the careful definition of common metadata fields supplemented by node specific metadata in support of publication venues such as WaterOneFlow web services for the CUAHSI ODM node. We have placed special emphasis on the semantic needs, i.e. built a support environment in which we attempt to provide a common set vocabularies in addition to some crosswalks between them. The EnvThs provides access to controlled vocabularies, taxonomies and ontologies widely used and recognized in geoscience/environmental informatics community. We use the Simple Knowledge Organization System (SKOS) for implementation deploying TemaTres, an open-source, web-based thesaurus management package and have also set up a SPARQL endpoint for programmatic access.

Keywords: Environmental Database, controlled vocabularies, SKOS, metadata, semantics

DATA STRUCTURE: ORGANIZING DATA

From previous experience on how to organize diverse data we concluded that data organization needs a broader context, such as a research project, a process, an event, as compared to just a geospatial reference such as a point or perhaps a region. We decided to adopt a project based organization scheme because extensive discussions with collaborators on this topic seemed to reveal a preference for this approach especially from individuals who are not familiar with the

details of cyber infrastructure solutions and also require minimum involvement and skill acquisition to submit data and subsequent search and retrieval. Hence every data item's primary affiliation is the project under which it was collected, as schematically shown in figure 1. Each data set can be a node specific set, i.e. a project could contain at least one set for each of the 6 nodes, but could also contain more. We also allow lateral use of Datasets; if a certain Dataset contains information that is used by another Project then it can be registered with that project also. In other words, our structure is not monotonic, but every data set can have multiple Project affiliations. A dataset in this context can be a collection from a one-time sampling or monitoring campaign which a defined start and end point in ti me, but also based on a network, in which sensors collect a multitude of diverse time series data that continue to get streamed into the servers as long as the network is operating. Lastly, we also found that this structure works well with the available (and required) metadata tags for each individual node, as we were able to map top level required metadata tags to our "Project" and "Dataset" layers thus ensuring homogeneity across the top level data descriptors. This in turn sets the base for permitting search across the nodes of the EDS.

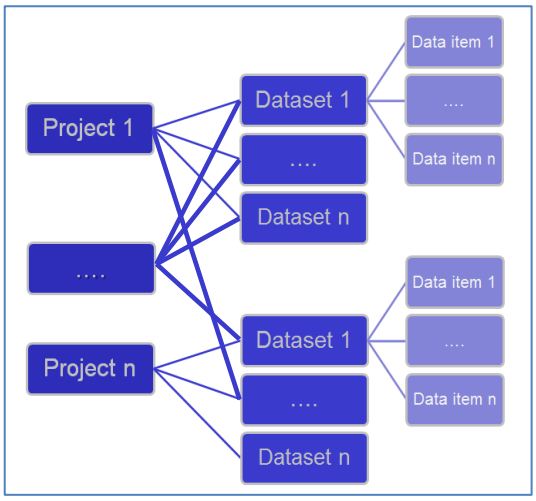

Figure 1. Data Organization in the EDS

COMMON DATA ELEMENTS

Common Data Elements (CDEs) are standardized terms for the collection and exchange of data. CDEs are metadata; they describe the *type of data being collected*, not the data itself. [2] Inside EDS, CDEs contribute to identifying detached items for data collection, improving consistent data collection under different context, and enhancing the ability of data discovering and sharing. Our six data types share some common attributes in addition to needing their individual sets of mandatory metadata. To this end we examined CUAHSI ODM 1.1 [3] and The Open Geospatial Consortium (OGC) WaterML 2.0 [4] conventions, which focus on time series data descriptions. We also investigated the 'Environmental Sampling, Analysis and Results Data Standards' (ESAR) series [5] and the Water Quality Data Elements (WQDE) [6] developed by the Environmental Data Standards Council (EDSC) and the National Water Quality Monitoring Council (NWQMC). These collections contain a large number of keywords and concepts that focus on in-situ sampled data descriptions. We also explored the a) ISO 19115 Metadata Standard for Geographic Information [7] which defines a total number of 31 code lists and enumerations for exchange of Geospatial data, b) Dataset Inventory Catalog Specification (DICS) [8] of THREDDS from which has a lot of lists used to describe modeling data. We extracted a total of 8 concepts as a starting point for the CDEs of EDS, as shown in figure 2, including *title, topic category, abstract, keywords,emporal coverage, spatial coverage, project*, and *contributor*. These are required at the beginning of the data submission process and are common to all submissions. The user is then directed towards node customized submission forms to complete the metadata requirements before the set can actually be submitted (we have

similar requirements in place for streamed data, which automatically annotates the incoming streams.

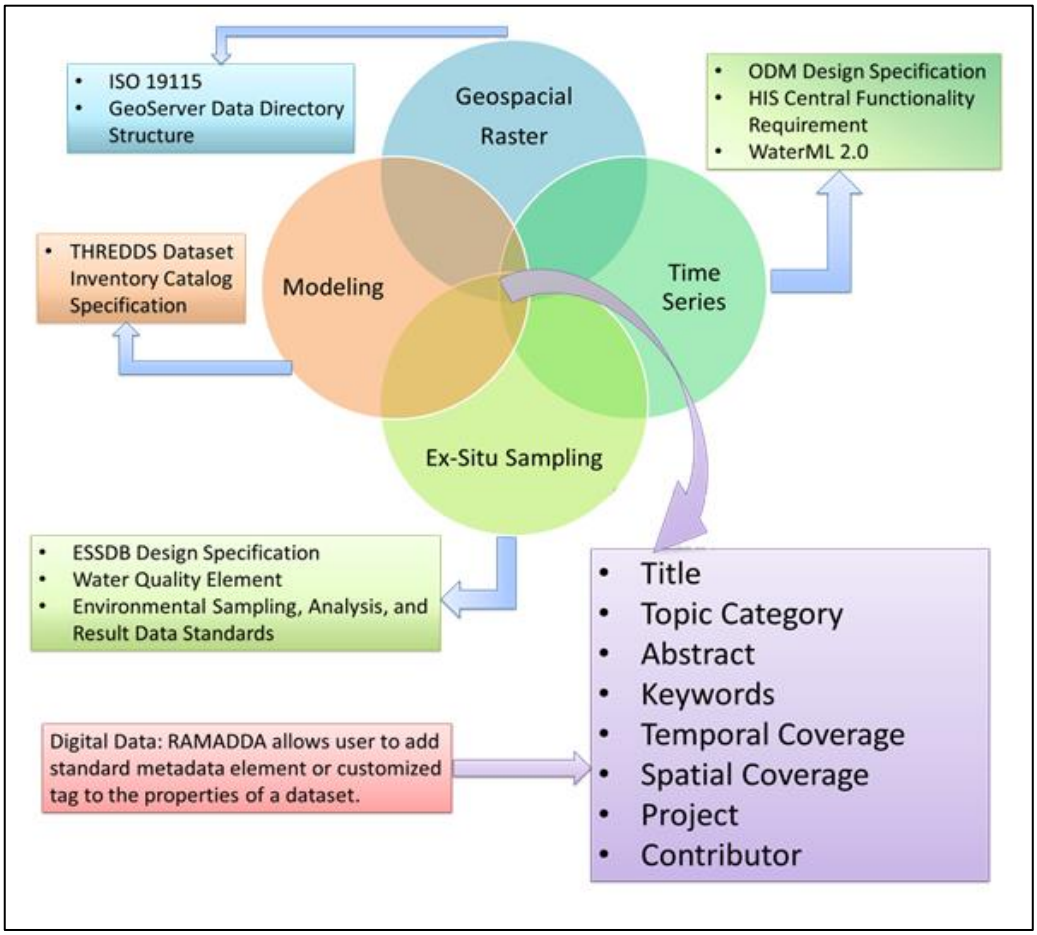

Figure 2. The common metadata element for the EDS

CONTROLLED VOCABULARY SERVICE APPLICATION

We developed a Controlled Vocabulary Service Application (CVSA) to bring some order and structure to the semantics in support of the EDS system. In doing so two aspects play an important role: the data model used and the vocabulary management software deployed. We adopted the SKOS [9] to represent the content of the EDS CVs. The SKOS represents a common data model endorsed by the World-Wide-Web Consortium (W3C) to support the use of knowledge organization systems (KOS) such as thesauri, classification schemes, subject heading systems and taxonomies within the framework of the Semantic Web. We selected TemaTres [10] as the management software to manage the content of the EDS CVs which is an open-source web application using the scripting language PHP and MySQL relational database management system, which also provides the capability of exporting vocabularies into RDF format (SKOS-core) in addition to XML format (Zthes, TopicMaps, MADS, Dublin Core, VDEX, BS 8723, SiteMap), SQL, and plain text. TemaTres was customized for use of the EDS CVs by integrating TemaTresView [11] and VisualVocabulary [12] to support tree view and visual representations of the CVs. We also installed the Open Source Edition of Virtuoso

Universal Server to provide a Web interface in support of SPARQL queries against the CVs. Restful web service provided by Tematres enables EDS system to retrieve vocabulary terms from our CVSA.

SUMMARY

Data structure, CDEs, and CVSA are fundamental components of our EDS application. We deployed a set of 6 dedicated nodes, one each for the 6 fundamental data types, to form the data store. We emphasized the use of open source software applications that have reached a high level of maturity and provided a common umbrella underneath which these nodes can be accessed for both deposit and retrieval. We have also developed an extensive controlled vocabulary system that has been published in a SKOS using TemaTres in addition to some visualization extensions. The EDS application has been deployed as an elementary and experimental prototype that can be accessed at http://endast.org.

ACKNOWLEDGEMENTS

We would like to acknowledge the National Science Foundation who has supported this work under grant numbers EAR0838307 and EAR0949196. We would also like to thank the City College of New York for their financial support for this project.

REFERENCE

[1] Michael Piasecki, Peng Ji, Conceptual Development of a Multi-Data-Type Enviornmental Data Store. 10th International Conferene on Hydroinformatics HIC 2012, Hamburg, Germany.

[2] NCI-Wiki CTEP, Common Data Elements, https://wiki.nci.nih.gov/display/caDSR/CTEP+Common+Data+Elements.

[3] ODM Controlled Vocabulary, The Consortium of Universities for the Advancement of Hydrologic Science, Inc. (CUAHSI), accessed Dec. (2013), http://his.cuahsi.org/mastercvreg/cv11.aspx.

[4] WaterML2.0 vocabularies, OGC, http://def.seegrid.csiro.au/sissvoc/ogc-def/resource?uri=http://www.opengis.net/def/waterml/2.0/.

[5] Environmental Data Standards Council (2006), Environmental sampling, analysis, and results data standards: Overview of component data standards, Stand. EX000001.1, Environ. Data Stand. Counc., U. S. Environ. Prot. Agency, Washington, D.C.(Available at http://www.envdatastandards.net/files/693_file_ESAR_Overview_01_06_2006__Final_.pdf).

[6] National Water Quality Monitoring Council (2006), Water quality data elements: A user guide, Tech. Rep. 3, Advis. Comm. on Water Inf., Washington, D. C. (Available at http://acwi.gov/methods/pubs/wdqe_pubs/wqde_trno3.pdf).

[7] ISO 19115 Metadata Standard for Geographic Information, http://www.iso.org/iso/catalogue_detail?csnumber=26020.

[8] Data Inventory Catalog Specification (version 1.0.2), UCAR, http://www.unidata.ucar.edu/projects/THREDDS/tech/catalog/v1.0.2/InvCatalogSpec.html.

[9] SKOS, Simple knowledge Organization System, accessed Sep. (2013),

http://www.w3.org/2004/02/skos/.
[10] TemaTres, accessed Sep.(2013), http://www.vocabularyserver.com/.
[11] TemaTresView, accessed Sep.(2013), http://www.r020.com.ar/tematres/wiki/doku.php?id=tematres:tematres_view.
[12] VisualVocabulary, accessed Sep.(2013), http://code.google.com/p/tematresvisualvocabulary/.