2013

A Justification for Semantic Training in Data Curation Frameworks Development

Xiaogang Ma
Rensselaer Polytechnic Institute, max7@rpi.edu

Benjamin D. Branch
Purdue University, bdbranch@gmail.com

Kristin Wegner
The GLOBE Program, kwegner@globe.gov

ED53C-0652 # A Justification for Semantic Training in Data Curation Frameworks Development

Xiaogang (Marshall) Ma (max7@rpi.edu)[1], Benjamin D. Branch (bbranch@purdue.edu)[2], Kristin Wegner (kwegner@globe.gov)[3]

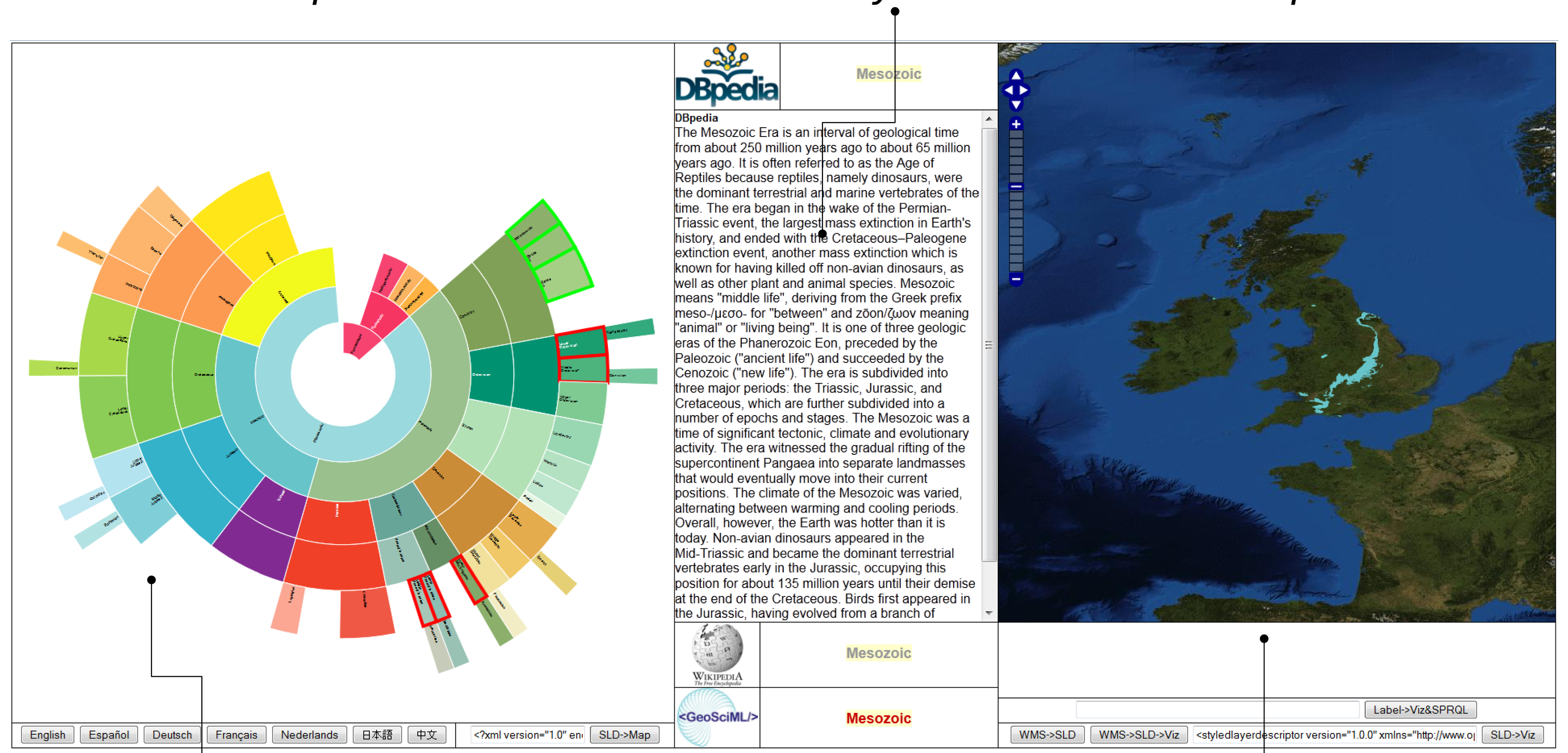
[1] Tetherless World Constellation, Rensselaer Polytechnic Institute, Troy, NY, USA; [2] GIS Department, Purdue University Libraries, West Lafayette, IN, USA; [3] The GLOBE Program, Boulder, CO, USA

ESIP 'Funding Friday' Award 2012

Background and Motivation

In the complex data curation activities involving proper data access, data use optimization and data rescue, opportunities exist where underlying skills in semantics may play a crucial role in data curation professionals ranging from data scientists, to informaticists, to librarians. Here, We provide a conceptualization of semantics use in the education data curation framework (EDCF) (**Fig. 1**) [1] under development by Purdue University and endorsed by the GLOBE program [2] for further development and application. Our work shows that a comprehensive data science training includes both spatial and non-spatial data, where both categories are promoted by standard efforts of organizations such as the Open Geospatial Consortium (OGC) and the World Wide Web Consortium (W3C), as well as organizations such as the Federation of Earth Science Information Partners (ESIP) that share knowledge and propagate best practices in applications.

DBpedia and GeoSciML vocabulary are sources for concept annotations



The functions of map legend and spatial feature filtration are re-developed with the visualization tool

An example of semantically enriched data service

Here we present an example that lowers the barrier of geologic knowledge for students by using semantic web technologies (**Fig. 2**) [3] . We first developed an geologic time ontology and deployed it to promote online geologic map services. Then we used the Linked Data resources (i.e., DBpedia and the GeoSciML vocabulary) for annotating multilingual geologic time terms and visualizations for presenting the results. Behind the user interface the geologic time ontology served as a basic reference for geologic time knowledge, and it controlled the reasoning of relationships between geologic time concepts, which is shown to users through the visualizations.

Obtaining an English term 'Triassic' from a map layer and retrieving Japanese annotations from Linked Data for the term

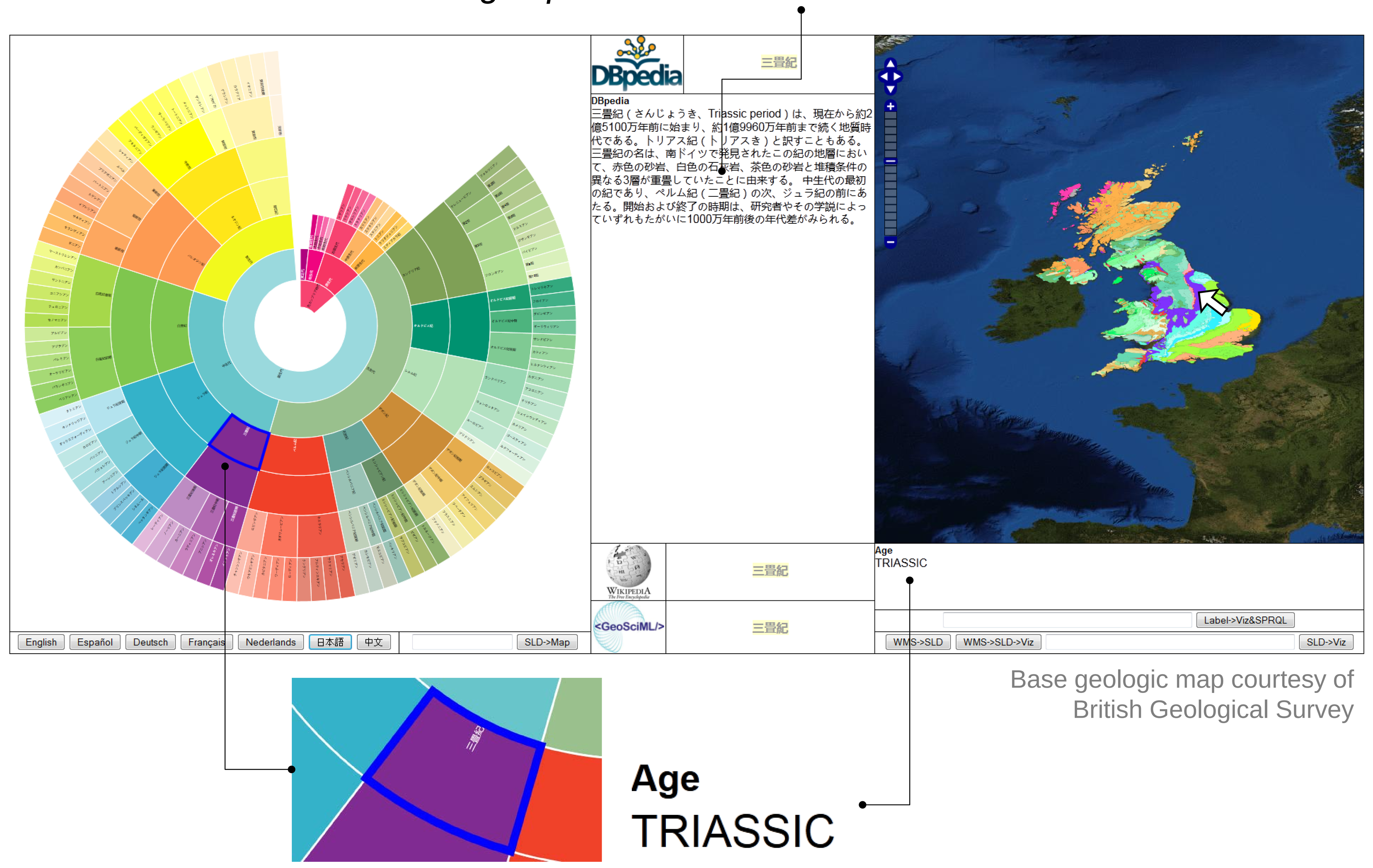


Fig. 2. Semantically enriched geoscience data services. *With minor to moderate mash-ups, semantic technologies can greatly improved the functionality of Earth and environmental science data services, such as the geological map service example shown here. Visualizations, on the other hand, can lower the barrier of domain knowledge to new comers, especially to the students.*

A Purdue Approach- conceptual data repository and outcomes diagram

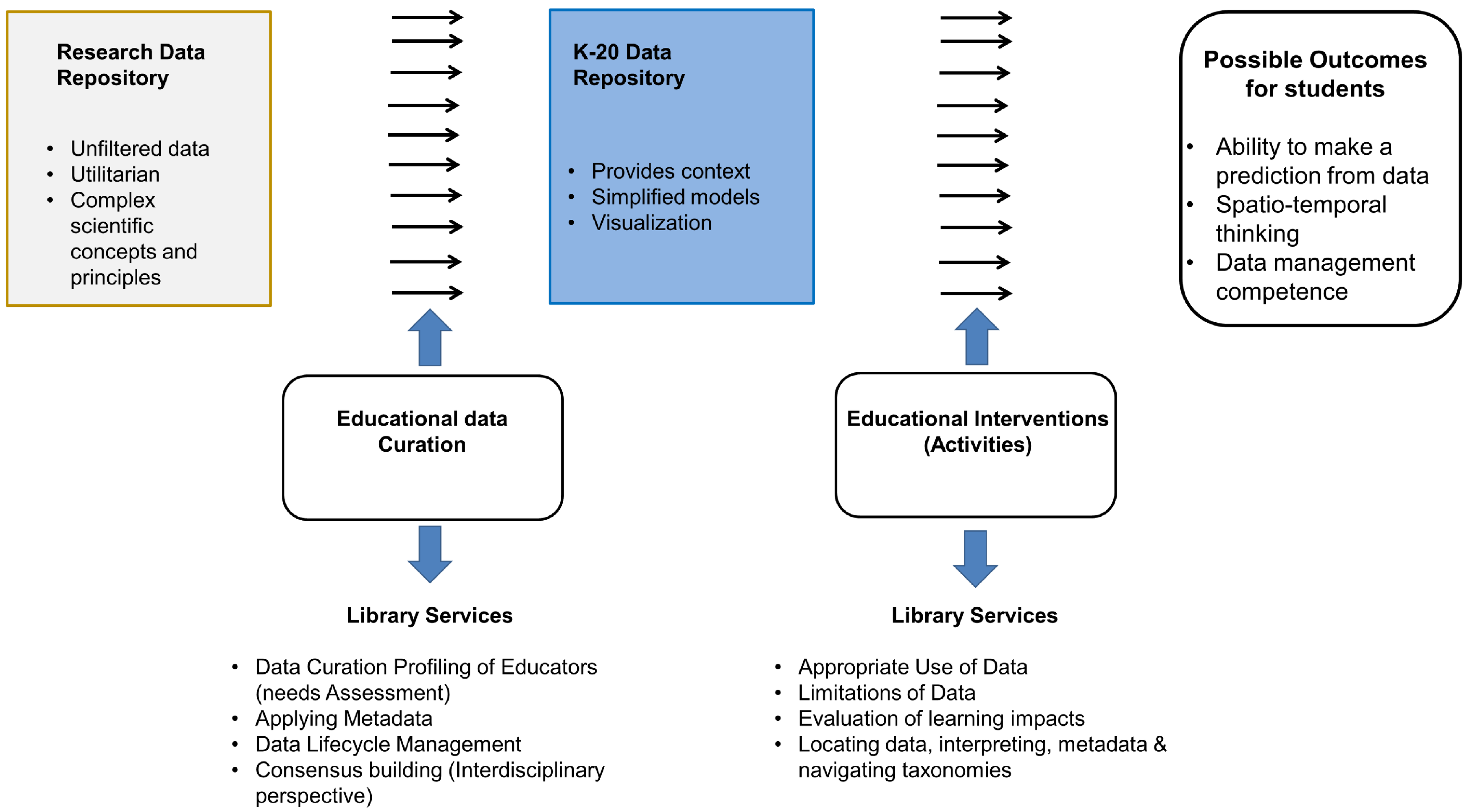


Fig. 1. Educational Data Curation Framework (EDCF) *– Is defined here as a Higher Education (HE) to K-12 knowledge transfer framework based upon the effective and interdisciplinary data science skills of future librarians working with all disciplines. The librarians conduct data curation properly at the HE level and have an ontological method to share the data in any possible place based or evidence based K-12 or primary education learning environment.*

Perspective

Outside the context of EDCF, semantics training may be same critical to such data scientists, informaticists or librarians in other types of data curation activity. Past works by the authors have suggested that such data science should augment an ontological literacy where data science may become sustainable as a discipline. As more datasets are being published as open data [4] and made linked to each other, i.e., in the Resource Description Framework (RDF) format, or at least their metadata are being published in such a way, vocabularies and ontologies of various domains are being created and used in the data management, such as the AGROVOC [5] for agriculture and the GCMD keywords [6] and CLEAN vocabulary [7] for climate sciences. The new generation of data scientist should be aware of those technologies and receive training where appropriate to incorporate those technologies into their reforming daily works.

References:

[1] Branch, B.D., Fosmire, M., 2012. The role of interdisciplinary GIS and data curation librarians in enhancing authentic scientific research in the classroom. American Geophysical Union 2012 Fall Meeting, San Francisco, CA, USA. Abstract# ED43A-0727
[2] http://www.globe.gov
[3] Ma, X., Carranza, E.J.M., Wu, C., van der Meer, F.D., 2012. Ontology-aided annotation, visualization, and generalization of geological time-scale information from online geological map services. Computers & Geosciences 40 107-119.
[4] http://www.whitehouse.gov/sites/default/files/omb/memoranda/2013/m-13-13.pdf
[5] http://aims.fao.org/standards/agrovoc
[6] http://gcmd.nasa.gov/learn/keyword_list.html
[7] http://cleanet.org/clean/about/climate_energy_.html

Acknowledgments:

* ESIP and TWC/RPI for funding the works shown in Fig. 2; * Nicole Kong, GIS Specialist & Assistant Professor, Purdue University Libraries; * Steven Smith, Earth, Atmospheric, and Planetary Sciences Outreach, Purdue University, A GLOBE Program partner