R2R+BCO-DMO – Linked Oceanographic Datasets

Adila Krisnadhi[1,2], Robert Arko[3], Suzanne Carbotte[3], Cynthia Chandler[4], Michelle Cheatham[1], Pascal Hitzler[1], Yingjie Hu[5], Krzysztof Janowicz[5], Peng Ji[3], Nazifa Karima[1], Adam Shepherd[4], and Peter Wiebe[4]

[1] Wright State University
[2] Faculty of Computer Science, Universitas Indonesia
[3] Lamont-Doherty Earth Observatory, Columbia University
[4] Woods Hole Oceanographic Institution
[5] University of California, Santa Barbara

Abstract. The Biological and Chemical Oceanography Data Management Office (BCO-DMO) and the Rolling Deck to Repository (R2R) program are two key data repositories for oceanographic research, supported by the U.S. National Science Foundation (NSF). R2R curates digital data and documentation generated by environmental sensor systems installed on vessels from the U.S. academic research fleet, with support from the NSF Oceanographic Technical Services and Arctic Research Logistics Programs. BCO-DMO human-curates and maintains data and metadata including biological, chemical, and physical measurements and results from projects funded by the NSF Biological Oceanography, Chemical Oceanography, and Antarctic Organisms & Ecosystems Programs. These two repositories have a strong connection, and document several thousand U.S. oceanographic research expeditions since the 1970's. Recently, R2R and BCO-DMO have made their metadata collections available as Linked Data, accessible via public SPARQL endpoints. In this paper, we report on these datasets.

1 Introduction

Researchers in the geosciences are challenged by the volume and heterogeneity of data types and formats, and the difficulty in discovering, accessing, and integrating data sets from multiple sources [2, 6]. At the same time, this diversity and heterogeneity is an unavoidable feature in a discipline that is so active and multi-faceted as the geosciences.

Geoscience researchers are therefore seeking methods and tools that allow them to more easily share, discover, access, and reuse data. Currently, a very important role to this end is played by large-scale data repositories, which warehouse data for redistribution and inspection. Each repository usually caters for a specialized subcommunity of researchers, and is highly specialized and focused on particular purposes.

In the meantime, the number of such repositories, which can be accessed on the World Wide Web, abounds. It thus comes as no surprise that they each come

with their own modes of access, visualizations, tools, data structures, etc. So, while access to relevant research data is now much easier *in principle*, diversity and heterogeneity continue to provide significant barriers to discovery and access.

At the same time, global issues such as climate change and deforestation, together with a growing understanding of the many interrelationships between different subdisciplines, impose the necessity to consider Earth as a single but very complex system. This drives the need to not only discover and access data, but also to integrate information accross fields and disciplines. This importance is witnessed, e.g., by the National Science Foundation's funding of the EarthCube program, which aims at providing "unprecedented data sharing" across the geosciences.[1]

Linked data, of course, provides a basic means to this end. Unfortunately, while the uptake of linked data in the earth sciences is growing, it also remains relatively slow. But as repository metadata begins to be published as linked data, it gathers momentum due to the additional opportunities provided by publishing in this shared format which decreases the barrier to reuse.

Another advantage of advancing linked data solutions for the geosciences emerges when considering the sociocultural benefits. For example, existing data compilations such as the Global Multi-Resolution Topography synthesis [8], Petrological Database [5], and Long Term Ecological Research Network [9] depend upon contributions from hundreds of individual stakeholders such as scientists and engineers on oceanographic cruises, geological surveys and mapping agencies, and students and postdocs working in laboratories. Providing attribution (credit) to contributors is imperative for the success of such syntheses. Publishing content as linked open data, including links to investigators and field expeditions, which, in turn, can be linked to journal articles and conference/award abstracts, will provide greater incentive to contributors. Combining linked data with greater semantic integration will not only facilitate connections between global/gridded synthesis data and expedition-based (point-, track-, time-series-) data, and make it easier for scientists to discover and access those data in a consistent manner for multi-disciplinary investigations; it will also generate enthusiasm among scientists to contribute their data.

In this paper, we present linked datasets providing content from the two key ocean science repositories in the U.S., The Biological and Chemical Oceanography Data Management Office (BCO-DMO) and the Rolling Deck to Repository (R2R) program. We will first discuss the specific relevance of these repositories and their datasets for their research fields (Section 2), then provide more details about the corresponding linked datasets and their availability (Section 3), before concluding (Section 4).

[1] `http://earthcube.org/`

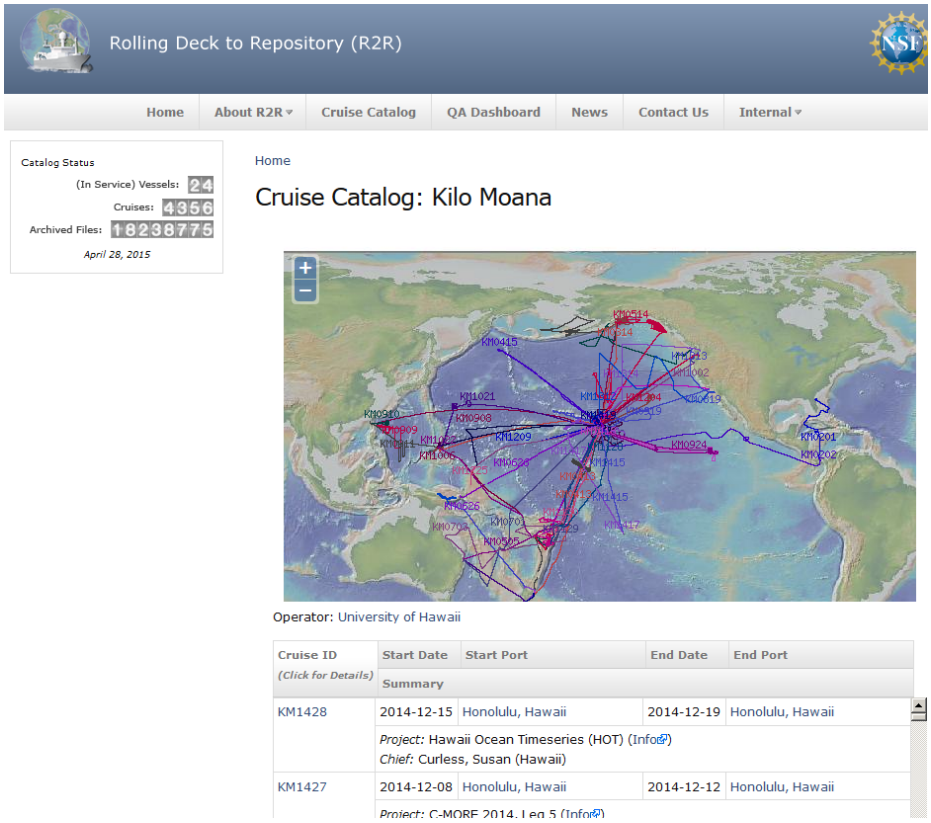

Fig. 1. R2R online user interface

2 Repository Description and Relevance

2.1 The R2R Program

With their global capability and diverse array of sensors, the U.S. academic research fleet is an essential mobile observing platform for ocean science. Data collected on every expedition are of high value, especially given the high costs and increasingly limited resources for ocean exploration. The Rolling Deck to Repository (R2R) program[2] is funded by NSF to provide stewardship of environmental sensor data routinely collected by the U.S. academic research fleet, working in close collaboration with the University-National Oceanographic Laboratory System (UNOLS) and the NOAA National Data Centers.

R2R maintains a catalog of vessels, instrument systems, expeditions, datasets, investigators, organizations, funding awards, cruise reports, and navigation tracks (see Figure 1) – every NSF-funded oceanographic cruise on a vessel in the academic fleet creates records in R2R. As such, R2R ensures preservation of and

[2] `http://www.rvdata.us/`

access to U.S. national oceanographic research data resources, and provides a central gateway through which data from oceanographic expeditions is routinely cataloged and securely transmitted to national long-term archives including the National Geophysical Data Center (NGDC) and National Oceanographic Data Center (NODC). R2R thus provides essential data documentation for each expedition, and tools to improve documentation of the wide array of shipboard data acquisition activities typical of modern expeditions.

R2R also conducts post-cruise quality assessment to document the quality of data as originally delivered from vessels and provides feedback to cruise operators regarding the data quality. The main objective is focused on identifying occurrences of *suspicious* data, and not to assess the scientific value of the data. That is, R2R aims to preserve the data and the accompanying metadata to capture as much as possible the orignal intent as they were collected or acquired during expedition. The quality assessment is realized through a series of (mostly) automated tests such as checking whether appropriate metadata exists, searching for possible errors in file formats, as well as collecting summaries of record-level testing of data. All of these are done without making changes to the original raw data files.

As of April 28, 2015, R2R hosts data from 24 in-service vessels, 4,356 cruises, and a total of 18,238,775 archived files. The R2R website has an average of over 60,000 page views per month.

2.2 BCO-DMO

The Biological and Chemical Oceanography Data Management Office (BCO-DMO)[3] was created to serve principal investigators funded by the NSF's Biological Oceanography, Chemical Oceanography and Antarctic Organisms & Ecosystems Programs as a facility where marine biogeochemical and ecological data and information developed in the course of scientific research can easily be disseminated, protected, and stored on short and intermediate time-frames. The Data Management Office also provides research scientists and others with the tools and systems necessary to work with marine biogeochemical and ecological data from heterogeneous sources with increased efficacy. To accomplish this, two data management offices were united in 2006 and enhanced to provide a venue for submission of electronic data and metadata and other information for open distribution via the World Wide Web. The BCO-DMO data system can accommodate many different types of data including biological, chemical, and physical measurements and results. The system provides access to the data (numbers, images, and/or documents) in a consistent manner, with sufficient metadata, so that others can make full use of these data for their own purposes. The existence of sufficient metadata enables the discovery and accurate reuse of data by more than just the initial investigators who collect and process the data. The BCO-DMO data system is not simply a catalog of data resources, but a system that takes full advantage of a MySQL database storing documentation (metadata)

[3] `http://bco-dmo.org/`

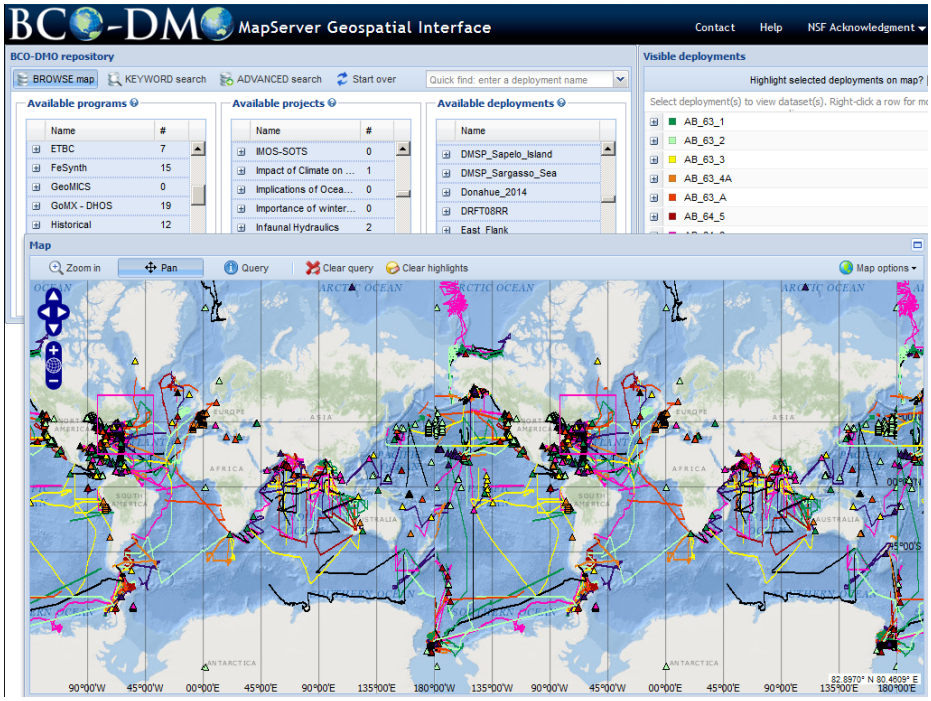

Fig. 2. BCO-DMO online map interface

for each data set, and a data management backend that allows data to reside at multiple sites (including the originating investigator's location if they wish).

The office manages existing and new data sets from individual scientific investigators and collaborative groups of investigators, and continues to make these available online. The office works with principal investigators and other data contributors on data quality control; maintains an inventory and program thesaurus of strictly defined field names; generates metadata Directory Interchange Format records required by federal agencies; ensures submission of data to national data centers; supports and encourages data synthesis by providing new, online, web-based display tools; and facilitates regional, national, and international data and information exchange. The data being served provide the scientific investigators with an opportunity to explore the complex and multifaceted data sets wherever they reside world-wide and to collaborate with colleagues in addressing pressing environmental questions, problems, and challenges. The BCO-DMO collection of data sets supports synthesis and modeling activities, reuse of oceanographic data for new research endeavors, availability of "real data" for teachers/students at school and college level to use in their classes, and provides decision-support field data for policy-relevant issues. Figure 2 shows a sample screen shot.

In terms of data quality, BCO-DMO employs an approach that is laregly people-intensive. Here, BCO-DMO provides data managers who work closely

with investigators to ensure sufficient metadata are collected and preserved to assist discovery, use, and reuse tasks. Collected metadata include information regarding design of experiments, instruments employed, as well as all the steps in processing field measurements into the final form of the data. Beyond the collection, data managers also coordinate closely with data contributors to decide how to organize and present the data in the best way possible. By employing this approach, BCO-DMO feels that higher quality data can be obtained and reused effectively.

As of April 28, 2015, BCO-DMO hosts 7,490 datasets including information about 1,799 researchers, 2,127 deployments, and 512 projects, that span the full range of oceanographic measurements from research cruises, timeseries sites, laboratory and mesocosm experiments, and synthesis and modeling projects. The BCO-DMO site typically has over 6,500 page views each month.

3 The Linked Datasets

3.1 R2R

The R2R linked dataset currently consists of over 530,000 triples, which are accessible via SPARQL Endpoint.[4] Machine-readable metadata is available at `http://data.rvdata.us/.well-known/void`. A Snorql interface is also provided[5] for exploring the SPARQL Endpoint, and an entry point URL is provided for Semantic Web browsers.[6] A navigable HTML view is also available.[7] The SPARQL endpoint is fed from the internal R2R database and is therefore up-to-date. Bulk download is possible at `http://www.rvdata.us/outgoing/lod/rvdata.us.20150430.ttl.gz`. R2R data are currently under Creative Commons CC BY-NC-SA 3.0 US license.

The RDF graph structure underlying the R2R linked dataset uses a set of interlinked ontology design patterns which are described elsewhere [3, 4]. A conceptual view on the schema can be found in Figure 3. Note that the triplification is done only on the metadata, and not down to each observation datum, which would require sheer amount of resources beyond the current capacity of R2R program. The ontology design patterns themselves are an ongoing recent outcome of the National Science Foundation's EarthCube program, more precisely of the GeoLink project[8] [10] and its precursor OceanLink [7]. They have been developed with ease of information integration in mind.

3.2 BCO-DMO

The BCO-DMO linked dataset[9] has machine-readable metadata accessible at `http://www.bco-dmo.org/.well-known/void`. The whole dataset currently con-

[4] `http://data.rvdata.us/sparql`
[5] `http://data.rvdata.us/snorql/`
[6] `http://data.rvdata.us/all`
[7] `http://data.rvdata.us/`
[8] `http://www.geolink.org/`
[9] `http://www.bco-dmo.org/linked-open-data`

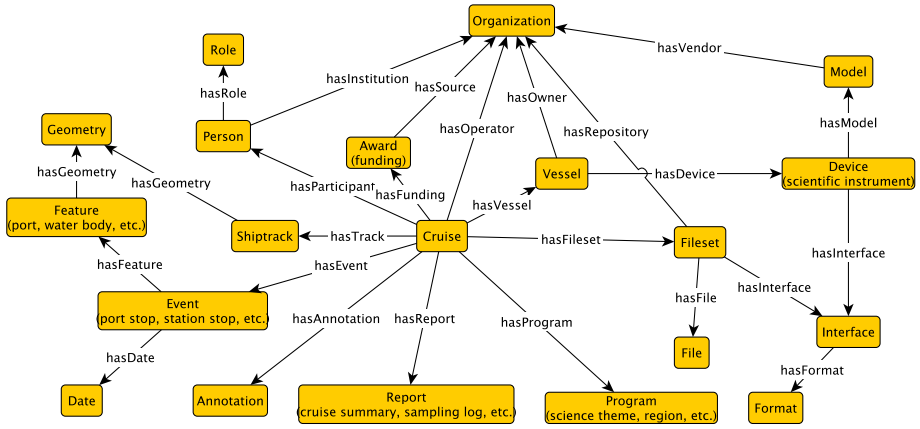

Fig. 3. R2R conceptual schema diagram

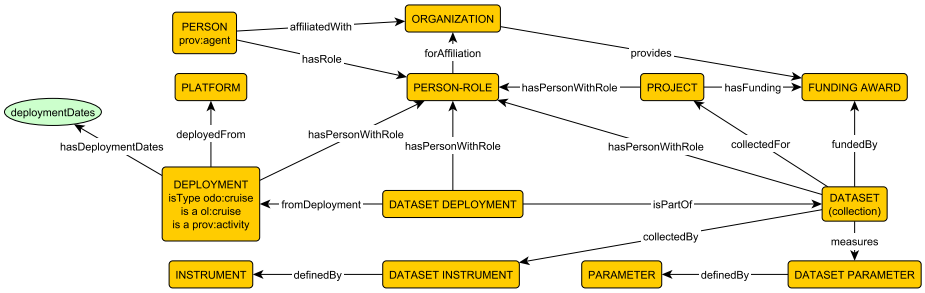

Fig. 4. BCO-DMO schema diagram

sists of over 2,170,000 triples. The triples are accessible via a SPARQL Endpoint and a Virtuoso SPARQL Browser[10] is provided for exploring the SPARQL Endpoint. The SPARQL endpoint is fed from the internal BCO-DMO database and is therefore up-to-date. Bulk download is also possible via the URIs pointed to by the `void:dataDump` property within the machine-readable metadata. BCO-DMO data are currently under Creative Commons CC BY-SA 3.0 license.

BCO-DMO uses a manually designed ontology for data organization, which was reported on in [1]. The schema diagram can be seen in Figure 4. Like in R2R, triples in BCO-DMO are essentially only on the metadata level, and not down to individual measurements. Meanwhile, for the purpose of better integration, not just with R2R, but also possibly with other data repositories in geo science, BCO-DMO provides additional triplification into the GeoLink design patterns, which are currently ongoing [4].

[10] `http://lod.bco-dmo.org/sparql`

3.3 The Overlap between R2R and BCO-DMO

The reader may suspect some overlaps exist between R2R and BCO-DMO, given that there are actually only dozens of oceanographic research vessels deployed for field observation, etc. The map-based interfaces also look similar. Indeed, there is a strong partnership between R2R and BCO-DMO, which makes linking their content between each other particularly attractive and potentially impactful. R2R housed data about the vessels, the route navigated during an expedition, as well as narrative description of activities performed during the expedition. It also hosted data obtained from on-board sensors and devices fixed to the vessels, such as those from CTD[11] instruments or multibeam sensors. On the other hand, data obtained from devices personally brought by the researchers (and thus are not fixed permanently to the vessels) are not kept by R2R, but rather by other repositories, particularly BCO-DMO. In this context, R2R and BCO-DMO are linked two each other via (meta)data about persons and they agree on oceanoraphic cruise identifiers. This linking is of high quality as both data repository maintainers closely cooperate to identify the overlap. For cruise identifiers in particular, there are only about a few dozens research vessels actively used for the U.S. oceanography research, and R2R essentially acts as the gateway of data from the whole fleet of vessels before data being deposited and catalogued in other long-term archives. As such, determining the mapping between the two datasets and checking the redundancy become relatively manageable. Furthermore, both linked datasets provide external links to DBpedia, more precisely they map affiliations (organizations), scientific instruments (devices), and research programs to DBpedia using skos:exactMatch links, these were discovered through string matching.

It is important to note that although BCO-DMO has information about cruises, it does not host the detailed navigation data and other kinds of data pertinent to the vessels of which the vessel operators are reposnsible – these are hosted by R2R. BCO-DMO is more focused on data from specific researchers who run research projects. This means that BCO-DMO would have more detailed data about observations and measurements made during a research expedition. In addition, BCO-DMO does not limit its operation solely on oceanographic data coming from expeditions aboard research vessels, but also those from deployments via other platforms, such as moorings, satellite, land-based platforms, or submarine-based platforms, although oceanographic data from vessel-based expeditions constitute significant chunk of the BCO-DMO repository.

4 Conclusion

As Semantic Web technologies are on the rise in applications, the publication of metadata as linked datasets by major geoscience data repositories is likely going to be a driver of future developments. As data becomes available as linked data, its reusability increases, and this includes the development of linked data based

[11] conductivity, temperature, and depth of the ocean

data discovery and access. In this paper, we have presented the linked datasets providing metadata for the two major oceanographic data repositories, R2R and BCO-DMO.

Besides the obvious potential these linked datasets have for leveraging Semantic Web technologies for the geosciences, these datasets also lend themselves to Semantic Web research, as they pose interesting and challenging problems while at the same time are "real" datasets, as opposed to the often artificial or academically produced benchmarks. For example, they provide an excellent playground for investigations into ontology matching due to the various degrees of overlap between sub-domains, widely different scales, and due to the fact that the utilization of spatio-temporal aspects will likely be critical. They also provide a realistic setting for co-reference resolution problems, solutions of which would have immediate beneficial benefit to the data repositories. Particularly interesting is the fact that, while the datasets are of significant size, they still center around a relatively clearly defined research community, thus certain variables can more easily be controled. Different ways to refer to places, e.g. via coordinates or gazetteer names, and different ways to refer to chemicals, e.g. by name or formula, etc. provide additional challenging dimensions for co-reference resolution research.

From a much wider perspective, of course, the development of Semantic Web methods and tools for on-the-fly integration of major geoscience data repositories would have immediate major impact on the work of geoscientists in practice. Providing linked data for some repositories – or even for most repositories – can only be a very small first step in this endeavour, which requires major advances in methods. Some EarthCube projects, among them the GeoLink project which the authors are part of, already pursue this vision.

Acknowledgement The presented work has been partially funded by the National Science Foundation under the award 1440202 "EarthCube Building Blocks: Collaborative Proposal: GeoLink-Leveraging Semantics and Linked Data for Data Sharing and Discovery in the Geosciences."

References

1. Chandler, C.L., Groman, R.C., Shepherd, A., Allison, M.D., Kinkade, D., Rauch, S., Wiebe, P.H., Glover, D.M.: Using Controlled Vocabularies and Semantics to Improve Ocean Data Discovery (Invited). AGU Fall Meeting Abstracts p. B5 (2013)
2. Heidorn, P.: Shedding light on the dark data in the long tail of science. Library Trends 57(2), 280–299 (2008)
3. Krisnadhi, A.A., Arko, R., Carbotte, S., Chandler, C., Cheatham, M., Finin, T., Hitzler, P., Janowicz, K., Narock, T., Raymond, L., Shepherd, A., Wiebe, P.: Ontology pattern modeling for cross-repository data integration in the ocean sciences: The oceanographic cruise example. In: Narock, T., Fox, P. (eds.) The Semantic Web in Earth and Space Science: Current Status and Future Directions. Studies on the Semantic Web, IOS Press (2015), to appear

4. Krisnadhi, A.A., Hu, Y., Janowicz, K., Hitzler, P., Arko, R., Carbotte, S., Chandler, C., Cheatham, M., Fils, D., Finin, T., Ji, P., Jones, M., Karima, N., Mickle, A., Narock, T., O'Brien, M., Raymond, L., Shepherd, A., Schildhauer, M., Wiebe, P.: The GeoLink modular Oceanography ontology. Submitted to ISWC 2015 (2015), available from `http://daselab.cs.wright.edu/topics/publications.html`
5. Lehnert, K., Su, Y., Langmuir, C., Sarbas, B., Nohl, U.: A global geochemical database structure for rocks. Geochemistry, Geophysics, Geosystems 1(5) (2000)
6. Malik, T., Foster, I.T.: Addressing data access needs of the long-tail distribution of geoscientists. In: 2012 IEEE International Geoscience and Remote Sensing Symposium, Munich, Germany, July 22-27, 2012. pp. 5348–5351. IEEE (2012)
7. Narock, T., Arko, R.A., Carbotte, S., Krisnadhi, A., Hitzler, P., Cheatham, M., Shepherd, A., Chandler, C., Raymond, L., Wiebe, P., Finin, T.W.: The OceanLink project. In: Lin, J., Pei, J., Hu, X., Chang, W., Nambiar, R., Aggarwal, C., Cercone, N., Honavar, V., Huan, J., Mobasher, B., Pyne, S. (eds.) 2014 IEEE International Conference on Big Data, Big Data 2014, Washington, DC, USA, October 27-30, 2014. pp. 14–21. IEEE (2014)
8. Ryan, W., Carbotte, S., Coplan, J., O'Hara, S., Melkonian, A., Arko, R., Weissel, R., Ferrini, V., Goodwillie, A., Nitsche, F., Bonczkowski, J., Zemsky, R.: Global Multi-Resolution Topography synthesis. Geochemistry, Geophysics, Geosystems 10(3) (2009)
9. Waide, R., Thomas, M.: Long-Term Ecological Research Network. In: Meyers, R.A. (ed.) Encyclopedia of Sustainability Science and Technology, pp. 6216–6240. Springer, Heidelberg (2012)
10. You, J.: Geoscientists aim to magnify specialized Web searching. Science 347(6217), 11 (2015)