LTER Databits
Information Management Newsletter of the Long Term Ecological Research Network

Fall 2014

Welcome to the Fall 2014 issue of Databits. The articles submitted for this issue cover a range of topics. A thought-provoking commentary presents a vision for archiving simulation model data and code - perhaps some of our readers will address current and emerging capabilities on that topic in a future issue. A guest contributor shares her experience in an exciting new program in Data Curation at the University of Illinois at Urbana-Champaign. LTER and DataONE combine forces to address a complex data synthesis challenge that is relevant to all LTER sites. Another feature article highlights a collaborative effort between scientists, artists, programmers, and educators to develop a visualization and sonification of the water cycle, driven by real-time sensor data. And finally, it wouldn't be Databits without a discussion of metadata. In this issue, we hear about new EML capabilities in the GCE Toolbox that improve interoperability with PASTA, and provide new features for workflow-driven data analysis. A second metadata article provides guidance on improving data discovery by internet search engines through the use of schema.org attributes in our web pages.

The balance of articles address data visualization, capturing spatial coordinates, and a useful R package - they are sure to provide helpful tips and inspiration.

DataBits continues as a semi-annual electronic publication of the Long Term Ecological Research Network. It is designed to provide a timely, online resource for research information managers and is supported by rotating co-editors.

Editors: Mary Martin (HBR), Wade Sheldon (GCE)

Featured Articles

Commentary

Good Tools And Programs

Good Reads

Featured Articles

WaterViz for Hubbard Brook: A New Water Cycle Visualization and Sonification Tool

edit

Lindsey Rustad (HBR, US Forest Service)

Investigators at long-term research sites, including many of the LTERs, are increasingly deploying a new generation of environmental sensors and wireless communications that are revolutionizing how we collect and share data about the natural world. These new digital devices allow for the collection and communication of gigabytes of environmental data in near real time, heralding a new era of science and discovery in the environmental sciences. The LTER network, together with partners from the US Forest Service, have taken a lead in developing online resources and publications which provide guidance for best practices for sensor networks and sensor data management (http://wiki.esipfed.org/index.php/EnviroSensing_Cluster; Campbell et al. 2013; Databits Spring 2014). We are now taking a lead in developing an entirely new medium for creatively communicating these near real-time data to the public: WaterViz for Hubbard Brook: A Water Cycle Visualization and Sonification Tool.

The WaterViz for Hubbard Brook lies at the nexus between the hydrologic sciences, the visual arts, music, education and graphic design. In a nutshell, hydrologic data are captured digitally from a small first order catchment at the USFS/LTER Hubbard Brook Experimental Forest in the White Mountains of New Hampshire using an array of environmental sensors. These data are transmitted to the internet and are used to drive a simple computer model that calculates all components of the water cycle for the catchment in real time. The complete set of measured and modeled data are then used to drive a flash visualization and sonification of the water cycle at Hubbard Brook, which are available to viewers and listeners worldwide on the Internet. The WaterViz provides a unique and novel approach that allows the viewer to intuit the dynamic inputs, outputs, and storage of water in this small, upland forested watershed as they are occurring and from anywhere in the world.

The visualization was developed as a collaboration between Forest Service and LTER scientists, a Finish forestry software developer, Simosol Oy, and artist, Xavier Cortada. The visualization animates different sections of Xavier Cortada's original painting of water flowing through the Hubbard Brook with moving particles. Each section of the painting represents a different component of the water cycle. The number of particles and the speed with which they move are directly proportional to, and driven by, the data.

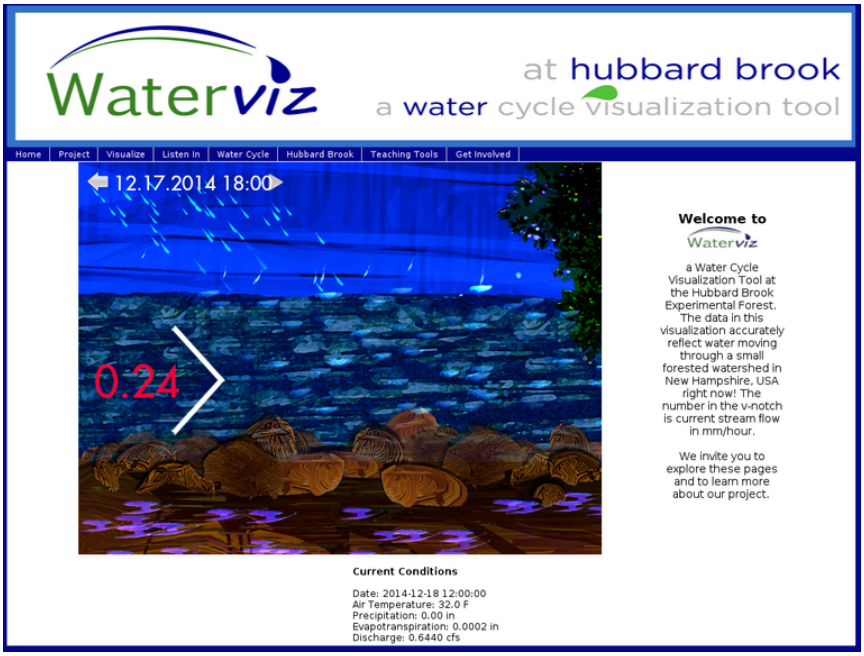
The sonification was developed as a collaboration between the same team of scientists and musician Marty Quinn, who runs the Design Rhythmics Sonification Research Lab. The sonification is the acoustic version of a data visualization. It connects the data to pitch, timing, and timber as played by different instruments, allowing the listener to 'hear' multiple lines of data in complex dimensions. We currently have two versions of the sonification: (1) Listen In, which allows you to hear five years of Hubbard Brook data sonified, and (2) Hubbard Brook Listen Live, which is an internet radio station playing live Hubbard Brook data. Both of the sonifications can be found on the Listen In page of the Waterviz website.

Educational applications are a priority, and we are working with a team of educators, modelers, and teachers to develop lesson plans to incorporate the Waterviz into middle- and high school curricula, using the principals of Universal Design for Learning (UDL). These principals emphasize the need to design flexible learning environments that can accommodate individual learning differences (Meyer et al. 2013). The premise is that some children (and adults) learn better visually, acoustically or even kinetically as compared to the more traditional cognitive reasoning pathways that dominate STEM curricula. The Waterviz,can be used to offer an alternative representation of the data that may help to engage and retain these students in the STEM disciplines, and more generally, communicate our data to a broader audience.

The Waterviz is an example of an emerging genre of online artistic visualizations of near real time data. Two other visualizations of near real time natural phenomena are The Carbon Tree and the Earth Visualization of Global Weather Conditions.

We invite you to view the Waterviz at: http://smartforests.org/waterviz, listen in to five years of data sonification or tune in to the live Hubbard Brook radio station at http://smartforests.org/waterviz/listenin.shtml. The station may be accessed using any Shoutcast capable Internet Radio Station app and searching for Hubbard Brook Forest Live. You can also learn more about the making of the Waterviz on our short video located on the WaterViz Project tab at: http://smartforests.org/waterviz/project.shtml.

The Waterviz is still under construction! We welcome and encourage feedback on the content and displays, and suggestions on creative collaborations and funding opportunities!



References:

Campbell, John L., Rustad, Lindsey E., Porter, John H., Taylor, Jeffrey R., Dereszynski, Ethan W., Shanley, James B., Gries, Corinna, Henshaw, Donald L., Martin, Mary E., Sheldon, Wade. M., Boose, Emery R., 2013. Quantity is nothing without quality: Automated QA/QC for streaming sensor networks. BioScience. 63(7): 574-585.

Meyer, A., Rose, D., and Gordan, D. 2014. Universal Design for Learning: Theory and Practice. Cast Professional Publishing, Cast Inc., Wakefield, MA. 234 pp.

Using the GCE Data Toolbox as an EML-compatible workflow engine for PASTA

edit

Wade Sheldon (GCE)

The GCE Data Toolbox for MATLAB was initially developed in 2000 to process, quality control and document environmental data collected at the then-new Georgia Coastal Ecosystems LTER site (Sheldon, 2001). Development of this software framework has continued steadily since then, adding graphical user interface dialogs (Sheldon, 2002), data indexing and search (Sheldon, 2005), web-based data mining (Sheldon, 2006; Sheldon, 2011b), dynamic QA/QC (Sheldon, 2008), and a growing suite of tools for automating data harvesting and publishing (Sheldon et al. 2013; Gries et al., 2013). We began distributing a compiled version of the toolbox to the public in 2002, and in 2010 we released the complete source code under an open source GPL license (Sheldon, 2011a). Today, the GCE Data Toolbox is used at multiple LTER sites and other research programs across the world for a wide variety of environmental data management tasks, and we are actively working to make it a more generalized tool for the scientific community (Chamblee et al., 2013).

The toolbox can be leveraged in many ways, but it has proven particularly useful for designing automated data processing, quality control and synthesis workflows (Sheldon et al., 2013; Cary and Chamblee, 2013; Gries et al., 2013). Key factors include broad data format support, a flexible metadata templating system, dynamic rule-based QA/QC, automated metadata generation and metadata-based semantic processing (fig.1). Consequently, the GCE Data Toolbox was one of the technologies chosen for a 2012 LTER NIS workshop convened to test the PASTA Framework for running analytical workflows (see http://im.lternet.edu/im_practices/data_management/nis_workflows). The lack of built-in support for EML metadata proved to be a significant barrier to fully utilizing this toolbox for PASTA workflows during the workshop; however, complete EML support has since been implemented. This article describes how the GCE Data Toolbox can now be used as a complete workflow engine for PASTA and other EML-compatible frameworks.

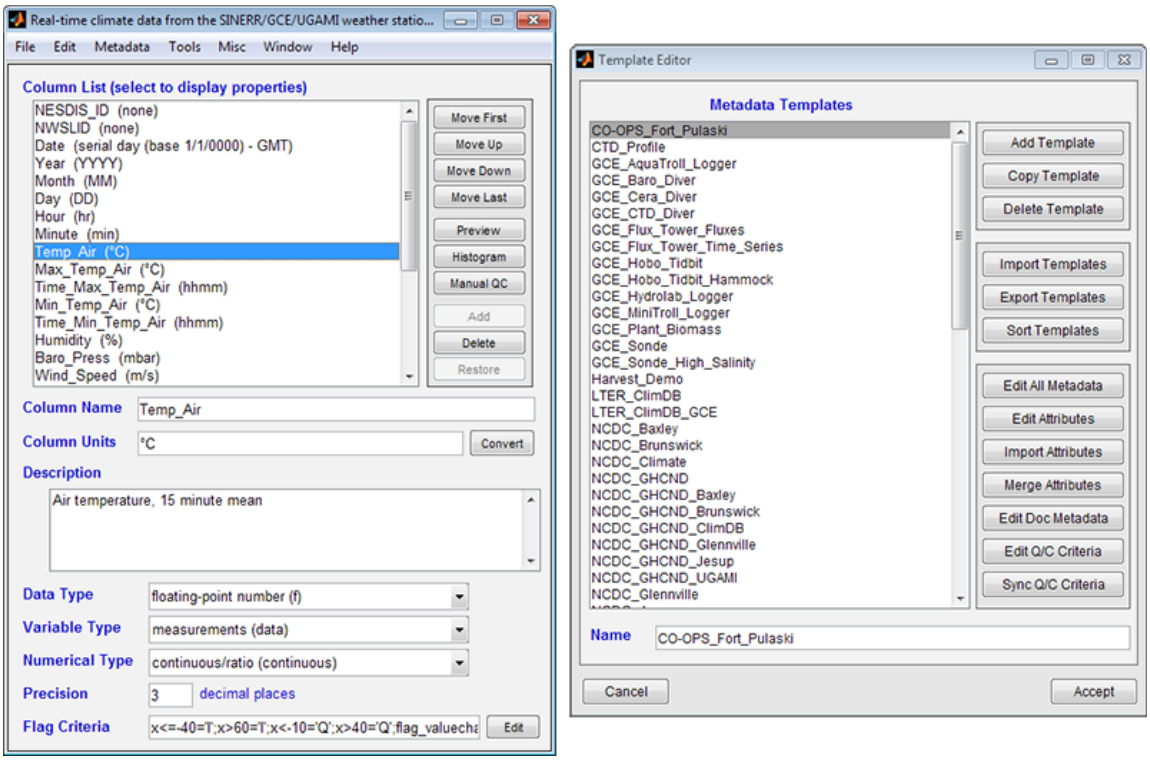


Figure 1. GCE Data Toolbox Data Set Editor and Metadata Template Editor GUI applications, with controls and options for managing metadata content and QA/QC rules to apply to harvested data.

EML-based Data Retrieval

The PASTA framework, as implemented to create the LTER Network Data Portal (https://portal.lternet.edu), is based on the EML data package concept and uses EML physical and attribute metadata to retrieve, validate and identify data objects that are stored in the system (Servilla et al., 2006). In order to download data from PASTA into the GCE Data Toolbox, code was developed to request an EML metadata document for a specified packageId using a PASTA web service. An XSLT stylesheet (EMLdataset2mfile.xsl) was then developed to transform the EML document into a native MATLAB program capable of downloading the described tabular data objects and parsing the data and metadata into MATLAB arrays, using the file structure and attribute metadata to generate appropriate command syntax. This approach was inspired by John Porter's PASTAprog, which transforms EML documents to generate R, SAS and SPSS code for retrieving and analyzing the data, and in fact this MATLAB stylesheet is now provided as an option for the PASTAprog web service at VCR and in the LTER Network Data Portal (see http://im.lternet.edu/im_practices/data_management/nis_workflows/PASTAprog). The generated program is saved as a MATLAB function m-file that can be run interactively or called in a workflow script. This function m-file is fully documented and can be archived along with the data, providing a means to re-download the same data in the future as well as useful provenance metadata for any workflows that leverage the data.

When the m-file function is called, a generic MATLAB data object (i.e. struct variable) is returned containing parsed metadata and data arrays organized into named fields. The data can be analyzed using standard MATLAB commands independently of the GCE Data Toolbox software; however, an import function was developed (eml2gce.m) to simplify transforming the parsed data and metadata into a toolbox-compatible data structure containing typed data, formatted documentation and attribute metadata, QA/QC rules and qualifier flags. Additional helper functions and GUI dialogs are also available to simplify mining data from PASTA and other EML repositories over the Internet, including the KNB Metacat and local site catalogs (fig. 2). Virtually any tabular text data that are properly described in EML (i.e. as dataTable entities and attributes) can now be retrieved into the GCE Data Toolbox with a single button press or workflow command, using only the structural metadata in EML to guide downloading, parsing and documenting of the data.

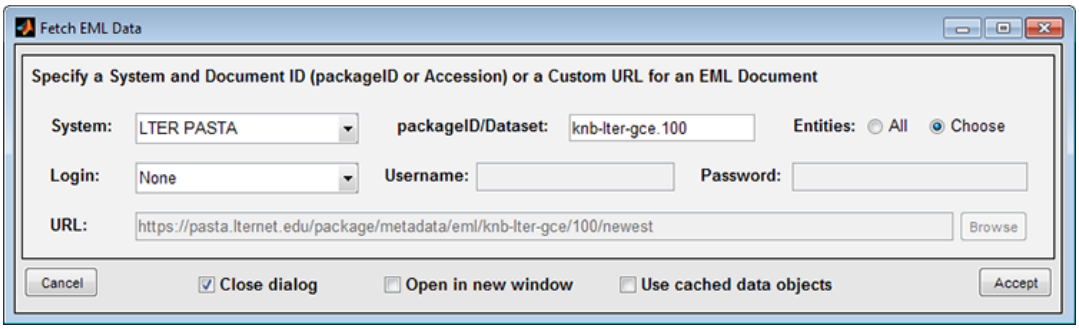


Figure 2. GCE Data Toolbox metadata content displayed in the Metadata Editor application, and styled as plain text, generic toolbox XML and EML.

EML Generation for Derived Data

Many types of workflows only need to read data from PASTA to create an analytical product, report or graph. Such workflows were the focus of the 2012 NIS workshop described above. However, workflows that synthesize data from one or more data sets in PASTA to create a derived PASTA data set, or that archive processed primary data in PASTA, are also potentially useful and were envisioned for PASTA since conception. In 2013 we attempted to create such synthesis

workflows during a follow-on PASTA workflows workshop at NTL (http://intranet2.lternet.edu/content/leveraging-pasta-and-eml-based-workflow-tools-lter-data-synthesis), but quickly ran into problems efficiently generating EML metadata for derived products we generated. Early plans for PASTA development included a "metadata factory" web service that could be used to programmatically generate EML in scripting environments. Unfortunately that service was scaled back to provide only provenance metadata fragments instead, requiring the workflow developer to generate the majority of EML content including the complex but critically-important attribute metadata. Manually authoring EML metadata using oXygen or Morpho proved too tedious and time-consuming to complete during the workshop, and was not recommend as a best practice for workflow development. A more automated approach for generating EML in workflows was clearly needed.

The GCE Data Toolbox is very adept at generating metadata for derived data sets during processing, by meshing metadata from source data sets and automatically creating attribute metadata for derived variables added by toolbox functions. The toolbox data model (i.e. GCE Data Structure; https://gce-svn.marsci.uga.edu/trac/GCE_Toolbox/wiki/DataModel) supports detailed, structured documentation and attribute metadata with fields based on the ESA's FLED report recommendations (Gross and Pake, 1995; Michener, 2000). A flexible metadata styling system is also available for transforming metadata into formatted text and XML documents (fig. 3). However, the toolbox data model pre-dates EML 2 by several years and intrinsically stores metadata content at lower granularity than EML requires (particularly personnel information), making cross-walking toolbox metadata to EML difficult and error-prone.

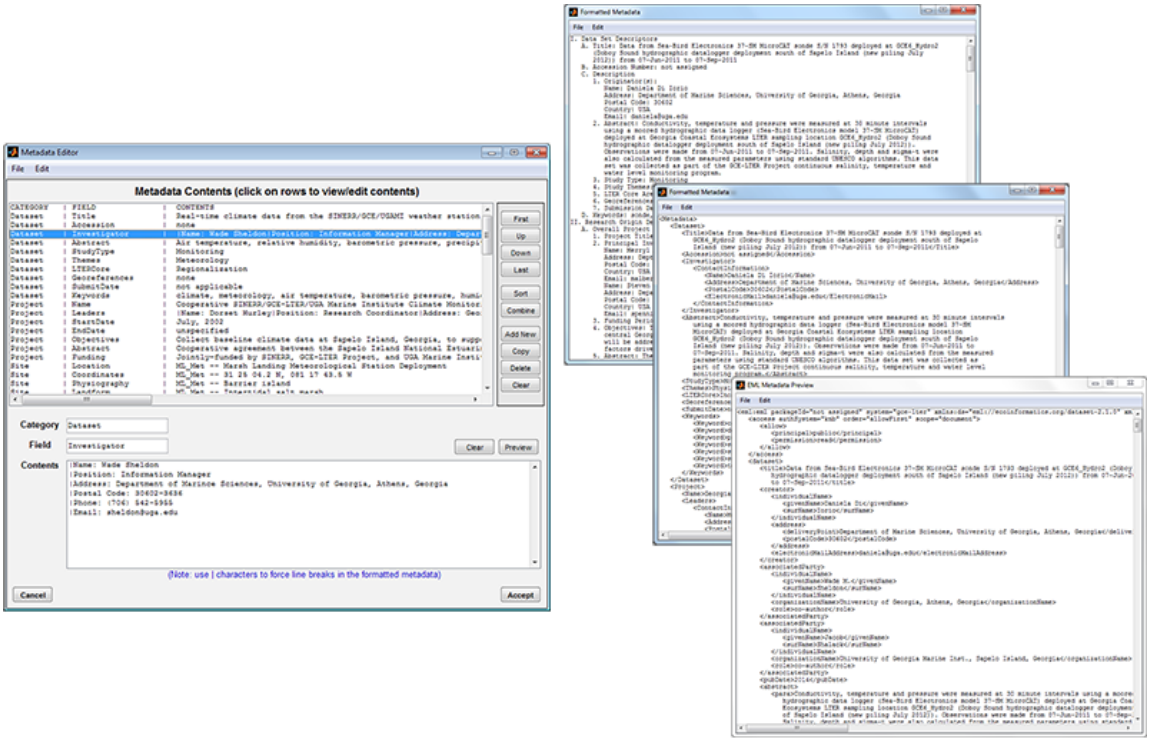


Figure 3. GCE Data Toolbox metadata content displayed in the Metadata Editor application, and styled as plain text, generic toolbox XML and EML.

In 2014 these difficulties were overcome, though, and support for producing fine-grained EML metadata was added to the GCE Data Toolbox. Native attribute descriptors (e.g. data type, variable type, number type, precision, code definitions, Q/C criteria) are automatically mapped to EML measurementScale equivalents, and units can be documented as custom units (complete with auto-generated STMML definitions in additionalMetadata), or can be mapped to user-specified standard and custom units managed in lookup-table data sets provided with toolbox downloads (i.e. EMLUnitDictionary.mat, EMLUnitMap.mat). Therefore once data are successfully loaded into the GCE Data Toolbox and described, EML with congruent data tables can be generated with no additional effort, removing a huge barrier for uploading data to PASTA. A GUI dialog is now provided with the toolbox for generating complete EML data packages, or just dataTable and attributeList XML fragments for inclusion in separately-generated EML depending on the desired use case (fig.4).

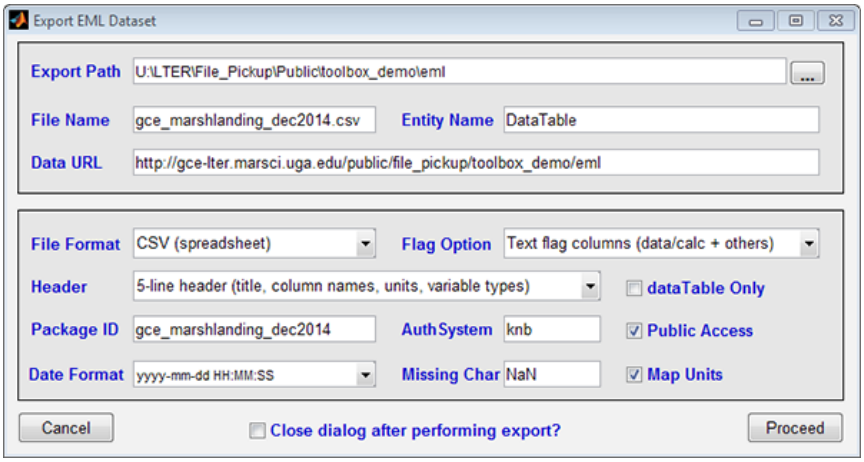


Figure 4. GUI dialog for generating formatted data files and corresponding EML metadata documents or fragments for uploading to PASTA or another metadata repository. Note that specific authorization metadata for non-public users cannot currently be specified using the GUI alone, but can be specified when calling the corresponding command-line function programmatically in a workflow.

Implementing Workflows

The GCE Data Toolbox provides a rich set of tools for processing raw data as well as importing and synthesizing existing data sets. For example, the toolbox can be used to programmatically mine data from the USGS National Water Information System, NOAA Global Historic Climate Network repository, NOAA Hydro-meteorological Automated Data System, LTER ClimDB/HydroDB, and DataTurbine servers directly over the Internet, providing a wealth of ready-to-use data for large-scale synthesis projects. Once data are imported, tools are provided for scaling and summarizing data by aggregation, binning and date-time resampling, as well as gap-filling, filtering and sub-setting data. Multiple data sets can also be integrated using database-style joins on key columns, as well as metadata-aware merges for concatenating related tables. For time-series data sets, overlapping date ranges in merged data can be removed automatically, and records can be padded to create a monotonic time series to simplify gap-filling and analysis. Unit conversion, data type transformation, date/time reformatting, geographic coordinate re-projection and other common data harmonization operations are also fully supported.

All of the operations described above can be performed using interactive GUI applications, but can also be scripted and run on a scheduled basis. The GCE Data Toolbox Wiki provides extensive documentation on getting started with this toolbox (https://gce-svn.marsci.uga.edu/trac/GCE_Toolbox/wiki/Documentation), as well as a quick-start guide to functions commonly used to build workflow scripts (https://gce-svn.marsci.uga.edu/trac/GCE_Toolbox/wiki/API). A graphical workflow-builder application is also envisioned for the toolbox in the future, but may require supplemental funding to implement.

Once EML-described data sets are generated from interactive or scripted workflows, the data objects need to be deposited in a web-accessible directory and the EML document uploaded to PASTA. To date we have only tested uploads using the LTER Network Data Portal web interface, but PASTA provides a comprehensive web service API that can be leveraged to script the entire evaluation and upload process for frequently-run workflows (https://pasta.lternet.edu/package/docs/api). Note that the POST and PUT commands necessary to upload EML documents require HTTPS and authentication, which are not supported by MATLAB's native HTTP functions (urlread, urlwrite). It is therefore necessary to call other programs via the MATLAB "system()" function to accomplish these steps. The simplest strategy is to install the cURL executable (http://curl.haxx.se/) on your system, which provides a rich set of command-line options for interacting with the PASTA API. The relevant cURL commands are also described in the PASTA API guide and draft workflow best practices guide (Gries et al, 2013b).

Conclusion

The primary rationale for the LTER Network's adoption of EML 2 as its metadata standard in 2003 was facilitating computer-mediated data analysis and integration. Unfortunately, the extensive effort required to upgrade legacy metadata content and management systems and the sheer complexity of this XML specification kept the focus on producing rather than using EML for much of the decade since. LTER sites can now produce EML metadata for the core data they archive, and PASTA has been implemented to provide stable access to version-controlled LTER EML documents and data, but the original goal for EML has proved elusive.

The NIS workflow workshops held in 2012 and 2013 demonstrated that effective workflows can indeed be built using EML-described data in PASTA and common research tools such as R, SAS, Kepler, and MATLAB. Now that EML with congruent, PASTA-compatible data files can be generated for synthetic data products automatically, hopefully we can return to this original goal and take full advantage of both EML and PASTA for LTER synthesis projects. The addition of EML support to the GCE Data Toolbox also provides LTER sites and other environmental programs with a practical solution for quality controlling and documenting streaming sensor data for archiving in an EML-compliant data repository (e.g. KNB Metacat, PASTA or another DataONE node) without the need to implement a full-fledged metadata management system (MMS), and provides new options for sites that have adopted a MMS like DEIMS or Metabase.

In other words, these improvements in the GCE Data Toolbox, along with the advancements made by the LTER Network Office NIS developers and the rest of the LTER information management community, go most of the way toward fulfilling our original vision from 2003 and all the way in terms of making workflow-driven analysis and archiving of streaming data into a reality.

Citations

Cary, R. and Chamblee, J. 2013. Coweeta LTER Upgrades Sensor Stations by Implementing the GCE Data Toolbox for Matlab to Stream Data. In: LTER Databits – Information management Newsletter of the Long Term Ecological Research Network: Spring 2013. LTER Network, Albuquerque, NM. (http://databits.lternet.edu/spring-2013)

Chamblee, J., Sheldon, W., Cary, R. 2013. GCE and CWT Host Successful Workshop to Demonstrate, Improve, and Promote the Adoption of the GCE Data Toolbox for MATLAB. In: LTER Databits – Information management Newsletter of the Long Term Ecological Research Network: Spring 2013. LTER Network, Albuquerque, NM. (http://databits.lternet.edu/spring-2013)

Gries, C., Sheldon, W.M. Jr., Fountain, T., Sebranek, C., Miller, M. and Tilak, S. 2013. Integrating Open Source Data Turbine with the GCE Data Toolbox for MATLAB. In: LTER Databits - Information Management Newsletter of the Long Term Ecological Research Network: Spring 2013. LTER Network, Albuquerque, NM. (http://databits.lternet.edu/spring-2013)

Gries, C., Porter, J., Ruddell, B., Servilla, M., Sheldon, W. and Walsh, J. 2013b. NIS data workflows best practices, ver. 0.2. Long Term Ecological Research Network, Albuquerque, NM. (http://im.lternet.edu/sites/im.lternet.edu/files/NISdataworkflowsbestpractices0.2.pdf)

Gross, Katherine L. and Catherine E. Pake. 1995. Final report of the Ecological Society of America Committee on the Future of Long-term Ecological Data (FLED). Volume I: Text of the Report. The Ecological Society of America, Washington, D.C.

Michener, William K. 2000. *Metadata*. Pages 92-116 in: Ecological Data - Design, Management and Processing. Michener, William K. and James W. Brunt, eds. Blackwell Science Ltd., Oxford, England.

Servilla, M., Brunt, J. San Gil, I., Costa, D. 2006. PASTA: A Network-level Architecture Design for Automating the Creation of Synthetic Products in the LTER Network. Ecological Informatics. (http://feon.wdfiles.com/local--files/start/LTERPASTADataModel.pdf)

Sheldon, W.M. Jr., Chamblee, J.F. and Cary, R. 2013. Automating Data Harvests with the GCE Data Toolbox. In: LTER Databits -Information Management Newsletter of the Long Term Ecological Research Network, Fall 2013 issue. LTER Network, Albuquerque, NM. (http://databits.lternet.edu/fall-2013)

Sheldon, W.M. Jr. 2011b. Mining Long-term Data from the Global Historical Climatology Network. In: LTER Databits - Information Management Newsletter of the Long Term Ecological Research Network, Fall 2011 issue. LTER Network, Albuquerque, NM. (http://databits.lternet.edu/fall-2011)

Sheldon, W.M. Jr. 2011a. Putting It Out There – Making the Transition to Open Source Software Development. In: LTER Databits - Information Management Newsletter of the Long Term Ecological Research Network, Spring 2011 issue. LTER Network, Albuquerque, New Mexico. (http://databits.lternet.edu/spring-2011)

Sheldon, W.M. Jr. 2008. Dynamic, Rule-based Quality Control Framework for Real-time Sensor Data. Pages 145-150 in: Gries, C. and Jones, M.B. (editors). Proceedings of the Environmental Information Management Conference 2008 (EIM 2008): Sensor Networks. Albuquerque, New Mexico. (http://gce-lter.marsci.uga.edu/public/files/pubs/wsheldon_dynamic_qc_eimc2008_final.pdf)

Sheldon, W.M. 2006. Mining and Integrating Data from ClimDB and USGS using the GCE Data Toolbox. In: LTER Databits - Information Management Newsletter of the Long Term Ecological Research Network, Spring 2006 issue. LTER Network, Albuquerque, NM. (http://databits.lternet.edu/spring-2006)

Sheldon, W.M. 2005. GCE Data Search Engine: A Client-side Application for Metadata-based Data Discovery and Integration. DataBits: an electronic newsletter for Information Managers, Spring 2005 issue. Long Term Ecological Research Network, Albuquerque, NM. (http://databits.lternet.edu/spring-2005)

Sheldon, W.M. 2002. GCE Data Toolbox for Matlab® --Platform-independent tools for metadata-driven semantic data processing and analysis. In: LTER Databits - Information Management Newsletter of the Long Term Ecological Research Network, Fall 2002 issue. LTER Network, Albuquerque, NM. (http://databits.lternet.edu/fall-2002)

Sheldon, W.M. 2001. A Standard for Creating Dynamic, Self-documenting Tabular Data Sets Using Matlab®. In: LTER Databits -Information Management Newsletter of the Long Term Ecological Research Network, Spring 2001 issue. LTER Network, Albuquerque, NM. (http://databits.lternet.edu/spring-2001)

DataONE to enable semantic searches for LTER NPP data

edit

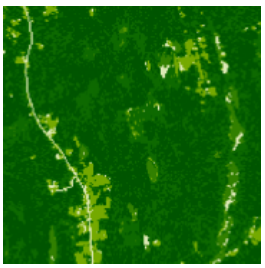
Margaret O'Brien (SBC)

Obstacles finding complex NPP data

The study of long term patterns in primary production is one of the core research areas at every LTER site, and a typically reported measurement is “net primary production” (NPP), e.g., the amount of new organic material produced during a time interval. Each LTER site’s measurements of NPP are determined locally and depend on the organisms being studied, e.g., their sizes, growth rates and community composition, and so methods vary widely in scope (organism, community, ecosystem), and scale (temporal and spatial). Methods also have different assumptions and limitations. It can be difficult to ascertain if any groups of NPP values are comparable without knowing these significant details. Scientists conducting synthesis projects using LTER data need a) to accurately find data sets containing NPP with the appropriate dimensions, and b) to learn enough about the methods in different studies to evaluate compatibility with their needs (Figure 1).

To enhance all data searches, the LTER information managers designed a SKOS-based controlled vocabulary. Through its use of “narrower terms” and “synonyms”, this has helped to refine search results in the Network catalog. But because the vocabulary’s scope is very broad, the NPP-related terms number fewer than 20 and there are essentially no terms related to field methods. A data collection as complex as LTER NPP data would be better served by an ontological system having strong semantic relationships and expressivity. The development of such a system is not possible with our current resources. Additionally, the issues surrounding description of NPP data are not unique to LTER, and the most complete, robust solution will be developed by a collaboration of data scientists and informatics specialists from many communities.

Figure 1. Methods for measuring NPP at LTER sites may have significant differences in temporal and spatial scales. Left, a chamber for measuring in situ NPP in a benthic algal community at the Santa Barbara Coastal LTER. Right, satellite image of NPP from the BigFoot site, Harvard Forest (image downloaded from ORNL DAAC).



Enter DataONE

The DataONE project began in 2009 and its large and diverse group of scientists and software engineers now coordinates hundreds of thousands of datasets from a diverse group of member nodes. LTER joined in 2012, contributing about 9000 records. Having recently entered Phase II, DataONE’s mission now includes solving specific problems for the earth science community. Its investigators are well aware of the difficulties that scientists face during data discovery, and plan cyberinfrastructure that will incorporate innovative and high-value features - among them, semantic technologies to enable precise data discovery and recall with measurement searches.To develop ontological solutions for data discovery, DataONE must begin with a “use case” - a sample problem which is constrained in scope, but complex enough to present a variety of potential obstacles, and with a large corpus of data having rich metadata. LTER’s diverse primary production data is an ideal case for developing a semantic search system.

Our range of measurements, biomes, and methodologies, comprising approximately 2500 datasets will ensure multiple benefits:

- DataONE is able to compare the effectiveness of an array of approaches
- LTER has a solution for one of its more complex data types
- The modeling patterns employed will be extensive enough to accommodate other future work in other scientific domains

Implementation plan

Broadly speaking, two major new components are needed for this semantic system: the ontology itself (also called a “knowledge model”), and the annotations to link datasets to concepts in the ontology. Many concomitant issues have been identified, particularly related to involvement with the the scientific community, including versioning, ownership, and design and usability of web-interfaces. All new technology is planned to build on pre-existing community practices for ontology and annotation formats and management. A working product is planned to be ready early in 2016, with a session planned for the 2015 LTER ASM to gather feedback on progress to date.

A group evaluated the current landscape of knowledge modeling for NPP data and ascertained that there were no pre-existing comprehensive models for this scientific domain. And so, DataONE’s work will serve not just its own and LTER’s needs, but the greater NPP community as well. DataONE has already dedicated a developer with an extensive background in coding for semantic systems. A computer science postdoc works on text and unit-matching algorithms. An LTER information manager acts as a coordinator, along with informatics scientists who have created other ontological frameworks. Under this group’s direction, two graduate students working at UC Santa Barbara/NCEAS will work on various aspects of assembly and annotation.

One important activity will be an “annotation experiment” to compare precision and recall for datasets with no annotation (e.g., the current text search), with manual annotation, fully automated annotation (based on text and unit matching algorithm), and semi-automated annotation (auto with additional choice/verification by users)

Extensions, relationship to other DataONE efforts

In addition to semantic search, DataONE also is implementing a system to enhance reproducibility by storing and indexing provenance trace information. The use case for this provenance work is the Multi-scale Synthesis and Terrestrial Model Intercomparison Project (MsTMIP), a comparison of carbon flux model results and observations, whose overall goal is to provide feedback to the terrestrial biospheric modeling community to improve the diagnosis and attribution of carbon

sources and sinks. MsTMIP's central measurement is Net Ecosystem Exchange (NEE), and MsTMIP has identified contributing measurements which are comparable to many LTER measurements, e.g., NPP.

MsTMIP needs differ from LTER's. Because the MsTMIP project is concerned with model results, its needs are primarily to track data provenance, and data discovery issues are secondary. However, scientists using MsTMIP data will need to: a) know which MsTMIP-model produced the dataset, and b) find appropriate benchmark data for model evaluation. Developing semantic structures for carbon flux should meet the discovery needs of both projects.

An initial focus on NPP data serves an important and timely need for the LTER, while setting the stage for future work. For example, continuing with the same projects, an extension could model and annotate the MsTMIP models themselves, which would also provide general examples for handling and modeling additional features of provenance. Another possibility is to improve discovery for data related to other carbon-cycling processes, e.g., oceanic carbonate system parameters for the study of ocean acidification. That work would almost certainly involve a broad community that intersected the LTER. Alternatively, we could tackle measurements in another LTER core research area. Nutrient flux is timely to consider because the Science Council plans to address this topic in upcoming synthesis projects. Semantic web technologies are an active area of development with high potential. But we know from our own experience that our data are too diverse to tackle at one time. We will be best served by breaking it into manageable chunks, and leveraging the work of those whose missions complement our own.

Google, Bing, Yahoo and your metadata

edit

Inigo San Gil (MCM), Stéphane Corlosquet (Aquia) and Adam Shepherd (ESIP - WHOI)

After years of suspense, the wait is over: The big three search engines have chosen a standard (aka specification) to provide information contributors with better mechanisms for describing information resources. The search engines improved the classification and sorting of information, resulting in a better experience when searching content on the web. When we say "content", we include datasets. This is the reason why data keepers should put attention to these particular advances by the main search engines and the reason we wrote this brief article.

The Google-Bing-Yahoo-Yandex chosen specifications reside in schema.org. The initiative was announced in June 2011, followed by workshops and early adopters (such as the White House). The first author of this article became aware of the Schema.org initiative during the last IM/ESIP meeting (ESIP, 2014). Here we expand on the Schema.org related topics covered at the ESIP Schema.org hack-a-thon session (Fils and Shepperd, 2014).

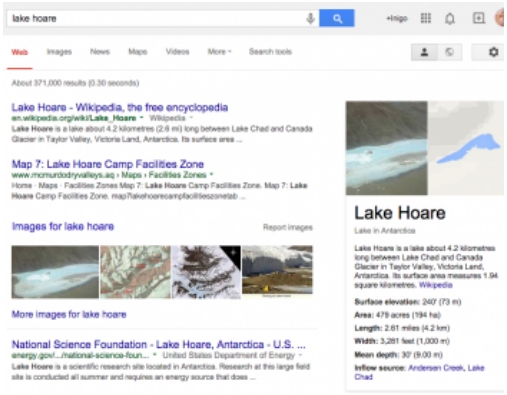
This article offers you a light view of the dataset specification at schema.org, a practical way to catch up with the schema.org specification, along with a motivation -- why would LTER comply with yet another metadata specification. The main merit of Schema.org adoption is to mitigate the failure in data discovery when the data seeker uses the main internet search engines.

Discovery: searching and finding.

Data Discovery is a hot topic for all LTER sites. What is the purpose behind Information Discovery science? Simply, the purpose of data discovery is to offer web visitors the best experience when searching for information resources, including data and research projects. A typical US-LTER site manages and serves thousands of unique information resource units including hundreds of scientific datasets. In this article, we refer to a dataset loosely as a group of contextualized scientific measurables. Structuring these information resources and making them easy to sort through is not a trivial task. At a much larger scale, the internet search engines are continuously improving their solutions to the similar but larger scale problem: sorting out information. In recent years, the race to offer the most relevant contextualized content has intensified. Specialized search results have been driven in part by the explosion of mobile devices -- download speeds and view port constraints forced search engines to be even more precise as, paradoxically, internet speed, device processor speed and screen size have decreased for the most popular devices used to utilize the Internet. Data set managers can take advantage of these new advances in search science. The specialization of the data-set annotations presented by schema.org represents one of the most relevant advances for scientific information managers.

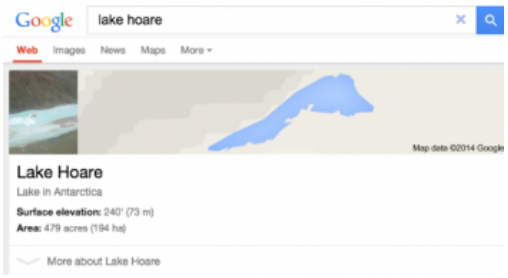
The *Internet of things*[1] will bring to you the Internet experience results that you are looking for by virtue of better indexing, cataloging and exposure to easily connectable services for data and information. You may count on the big search engines to make this happen; after all, those companies have all the ingredients to perfect the art and science of information management.

For over a year or more now, you surely noticed the mobile-friendly rich snippets that google offers as a result of a search. Locations, corporations, all sort of things you look for appear now in a brief vignette with relevant info, which may include a title, short description, geo-location and temporal relevance. Depending on the information sought, the bits highlighted vary. We experimented with these new features. We looked for "Lake Hoare" using Google. We also repeated the search using Bing and Yahoo with similar results. Here we present information for the Google results. The *Lake Hoare* search at the time of writing showed a rich snippet on the right side of a Chrome browser using a Windows Desktop computer.



The results offer a photo, a map and a brief summary of relevant data such as surface elevation, area, length, width, mean depth and the inflow sources. There is also attribution information (credit to the sources).

Similarly, the same google search using Safari on a 64Gb iPod touch 5G yields an even more brief rich snippet, which includes the photo, mini-map and 2 informational tokens (Elevation surface and Area), plus a "Read more" tab. It is also worthwhile to mention that the classic results come after this *rich snippet*.



Likewise, a dataset search on google, if annotated appropriately, may yield a similar snippet, perhaps a representative graph and a few key information placeholders, such as the where (a map) and the when, plus an attribution link.

What is Schema.org?

Schema.org describes itself as follows: "[] provides a collection of schemas that webmasters can use to markup HTML pages in ways recognized by major search providers, and that can also be used for structured data interoperability (e.g. in JSON). Search engines including Bing, Google, Yahoo! and Yandex rely on this markup to improve the display of search results, making it easier for people to find the right Web pages." The charge is to provide a unified structure to accomodate descriptions of all things, and use them for better discoverability. The interoperability dimension will not be discussed here.

Google, Yahoo and Microsoft's choice of vehicle to annotate datasets is not EML, the Biological Data Profile, the Dublin Core, the Darwin Core or any of the 1911* family of standards promoted by the International Standards Organizations. The good news is the actual Schema.org concepts and implementations are synergistic to those adopted by LTER over a decade ago; however, the new conventions and their particular technical implementations are different enough that you may have to deal with a fresh adoption process. Structurally, the markup is not encoded in XML, but rather simple HTML attributes, with proper namespaces. The juxtaposition of defined elements is far from 100% relative to LTER's usage of EML. Details about the concept overlap and a complete detailed discussion of the mappings between the current used standards and Schema.org is beyond the scope of this article. However, the authors encourage the adopters to revise our current implementation and report missmatches and possible new mappings.

Schema.org technical details

The Schema.org conforms to a hierarchical structure. At the top of the hierarchy, there are two elements and concepts: The Thing and the DataType. The most relevant element for the purpose of this article is the DataSet which is under CreativeWork (Note: do not get trapped by the term's semantic connotations - for now please see it mainly as hierarchical plausible placement). The element CreativeWork has the following differentiated information placeholders branching out:

`about, accessibilityAPI, accessibilityControl, accessibilityFeature, accessibilityHazard, accountablePerson, aggregateRating, alternativeHeadline, associatedMedia, audience, audio, author, award, awards, citation, comment, commentCount, contentLocation, contentRating, contributor, copyrightHolder, copyrightYear, creator, dateCreated, dateModified, datePublished, discussionUrl, editor, educationalAlignment, educationalUse, encoding, encodings, genre, hasPart, headline, inLanguage, interactionCount, interactivityType, isBasedOnUrl, isFamilyFriendly, isPartOf, keywords, learningResourceType, mentions, offers, provider, publisher, publishingPrinciples, review, reviews, sourceOrganization, text, thumbnailUrl, timeRequired, typicalAgeRange, version, video.`

The placeholders above contain familiar items, such as *creator, publisher, datePublished* and *keywords*. There are also a few that may have a match with placeholders we have been using to describe data sets. For example, *alternativeHeadline, citation, version, inLaguage* and *contentLocation*. Many of these terms are also parents of other terms in this hierarchy of things, such as "creator" which is of the type "person".

Many of the *CreativeWork* concepts that seem to be encapsulated in placeholders such as *hasPart, isPartOf, mentions, comments, typicalAgeRange, interactionCount, contentRating* and *interactivityType* are not considered in our EML schema descriptions. Perhaps some of them deserve a second look, specially those that indicate relationships, such as isPartOf and hasPart. Relations between datasets and information is one of the weakest parts of EML. The EML relational potential exists, but in practice it never translates into discoverability through relationships. For example, our network lumping techniques resulted in overly complex hierarchies.

Finally, the Schema.org *DataSet* adds a few data-set specific properties to the properties inherited from the *CreativeWork* parent category:

- catalog - A data catalog which contains a dataset
- distribution - A url that points to the data-resource
- spatial - The geo properties of the data
- temporal - A date range that characterizes the data-set

These four are well known by anybody documenting datasets in ecology using ISO, EML or BDP, and you should not have issues populating these schema-org compliant html attributes accordingly. The Schema.org can be extended to meet needs that may go beyond discoverability; however, the main goal for the Googles of the world is just to find resources using their search engine services. An example of the extension exercise is explained by Barker (2014) using concepts and placeholders borrowed from the Dublin Core.

Schema.org and DEIMS: An implementation case.

One of the advantages of adopting DEIMS is that you ride along a community of developers that can do. Our colleagues at ESIP were already developing an extension to accomodate the mappings from our database to the HTML rendering of schema.org categories. The latest version of DEIMS is schema.org compliant out of the box.

What if you are using Drupal, but not DEIMS? At the time of writing, some LTER sites (e.g., Virginia Coast Reserve, Coweeta, California Current and Palmer LTERs) use Drupal in their hybrid information management systems. Perhaps these sites can still leverage the Drupal work. I will describe what we did for DEIMS so perhaps you can re-use some of these steps. First, we installed the Drupal Schema.org module. The install process involved issuing 2 commands on the DEIMS server's bash shell:

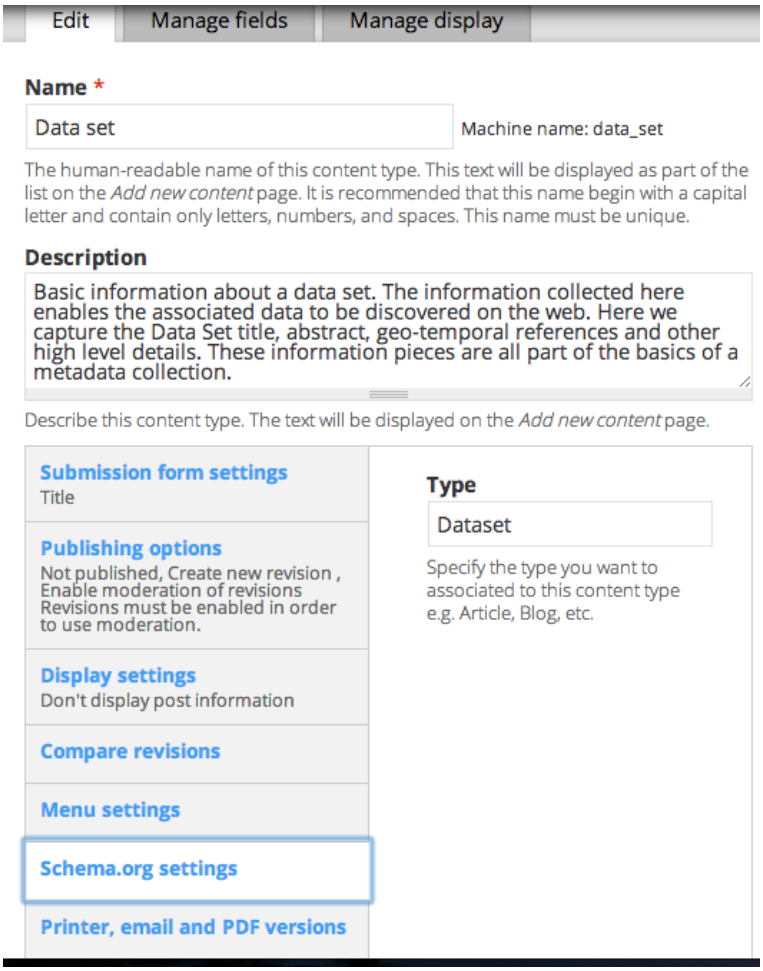
% drush dl schemaorg and

% drush en schemaorg schemaorg_ui

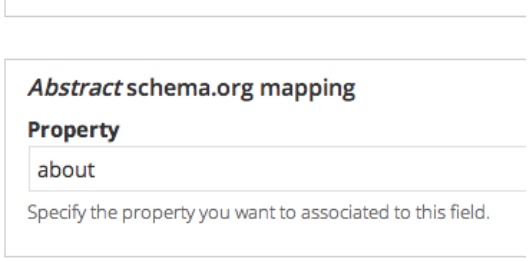
You can do this first step the traditional Drupal way -- installing a module like any other Drupal contrib module without using drush.

The next step involves configuring the newly installed schemaorg module. This step can be broken down into two general sub-steps.

The first configuration sub-step: Using your favorite browser, visit the edit dataset content type (the URL tail will look like admin/structure/types/manage/data_set). In the vertical menu, locate the schemaorg settings, and using the autocomplete associate the DEIMS dataset content type with schemaorg's 'Dataset' type. See figure below.



The second configuration step is to map all the mappable fields from the DEIMS Dataset content type to Schema.org's Dataset specification placeholders. For example, the "Abstract" maps to Schema.org's "About". For our mapping, we clicked on the manage fields tab for the Dataset content type, and looked for the "Abstract" field in the new form. We clicked on edit for that abstract field, and then at the bottom of the abstract field configuration form, we found a schema.org mapping autocomplete text field, which we used to find "About" and map the "Abstract". See figure below.



We repeated this DEIMS-field to Schema.org Dataset property mapping process for all the fields we thought were mappable. That was all there was to it: the rendered HTML produces the markup that optimizes the job of big search engine robots.

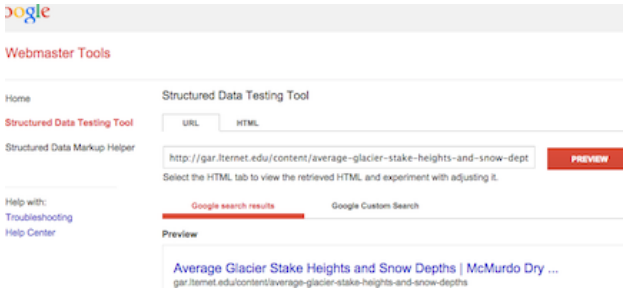
If you are using XSLT to render your Dataset pages on webpages, I would suggest you add XLST directives to produce Schema.org compliant attributes to your resulting HTML. THe process is undoubtedly more laborious, but not too daunting. Rendering content as HTML attributes is one thing the XSLT may do reasonably well.

Testing Schema.org implementation

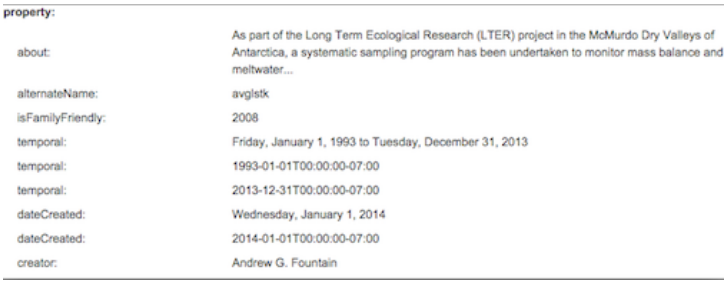
Once you have worked your way to big SEO rich snippet compliance, it may be time to see your results. Google provides this testing tool to examine whether your markup is producing the results as advertised: http://www.google.com/webmasters/tools/richsnippets

The tool would allow you to write a url, and receive a redux-report of what google is able to parse out.

The landing page of the tool also loads the results, which seems logical and less entropic. Let's test McMurdo's Glaciers data!



The result of the test is good. The data set title is highlighted, just as expected. How about the rest of the metadata? Many attributes were parsed correctly, and it even does well with dates, the *bête noire* of metadata. Here is a screenshot of some of the attributes, as parsed by google:



property:	
about:	As part of the Long Term Ecological Research (LTER) project in the McMurdo Dry Valleys of Antarctica, a systematic sampling program has been undertaken to monitor mass balance and meltwater...
alternateName:	avglstk
isFamilyFriendly:	2008
temporal:	Friday, January 1, 1993 to Tuesday, December 31, 2013
temporal:	1993-01-01T00:00:00-07:00
temporal:	2013-12-31T00:00:00-07:00
dateCreated:	Wednesday, January 1, 2014
dateCreated:	2014-01-01T00:00:00-07:00
creator:	Andrew G. Fountain

There is some work to do, but mostly the test gives satisfactory results. You may also wonder how many days it took us to implement this. The answer would not be of much guidance to many: I just applied the schemaorg module to DEIMS, hooray for Steph Colorsquet and others from the Drupal community.

Discussions: It is Google, Microsoft, Yahoo! and Yandex.

Before concluding the article, we would like to encourage you to make an effort to mark your datasets according to the rules set forth by the big three search engines and Yandex. There are many scientists that will not use google to search for a dataset, but some scientists will, and undoubtedly many internet users will bump into rich snippets featuring your site data sets once these have been properly marked up. Another good reason to adopt these rules is the derived tools that will likely be built to work with the Google, Microsoft and Yahoo specifications. We will want to leverage some of these tools, and chances are these will be developed at a faster pace than those supported by small groups of poorly funded initiatives.

Adoption of Schema.org may boost data discoverability, but it is also about taking advantage of the potential that schema.org and the companies behind it offer, and having the opportunity to help shape what interoperabilty might be in the era of the Internet of Things.

At the time of writing we could not offer a comparative analysis of the user experience, as the novel implementation has not yet gathered sufficient statistical convergence to discuss any possible improvements.

Footnotes

1 The internet of things is expected to offer advanced connectivity of devices, systems, and services that goes beyond machine-to-machine communications (M2M) and covers a variety of protocols, domains, and applications. [Wikipedia - the internet of things]

Citations

Fils, Dough and Shepherd, Adam. Schema.org Hack-A-Thon. ESIP 2014 Summer meeting, Frisco, CO. Resource at: http://commons.esipfed.org/node/2557

Schema.org Hack-A-Thon

Schema.org. Here is the thing. https://schema.org/Thing

The internet of things. About 43,300,000 results as of Nov. 2014. Google.

Web Schemas http://www.w3.org/wiki/WebSchemas

Corlosquet, Stéphane. 2011. Drupal contrib module schemaorg. http://www.drupal.org/project/schemaorg

Barker, Phil and Campbell, Lorna M. The Learning Resource Metadata Initiative, describing learning resources with schema.org, and more? Nov, 2014. Webinar. Resource at http://bit.ly/1pKiCUj

Becoming an Information Professional: A Student Experience with UIUC MLIS Program's Data Curation Specialization

edit

Chung-Yi Hou, Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign

As the volume, format types, and sources for data increase rapidly with the invention and improvement of new scientific instruments, the ability to manage and curate data is becoming more important as well. The skills and knowledge required to provide stewardship for digital data are especially crucial. This is because the rate of digital data generation and usage significantly outpaces the number of trained informational professionals who are available to support the varying data requirements from all the different research domains. In fact, many major studies have found and reported the need to provide a sustainable framework and policies that will allow on-going programs to be implemented to develop and train skilled information professionals. For example, the Research Information Network (RIN) and the British Library proposed that policy-makers need to help with "producing appropriate, effective and sustainable models for training and

careers in managing data" (Research Information Network and the British Library, 2009, p. 52). Likewise, the European Commission published a report in 2010 emphasizing the need for data scientists and their expertise by placing the following call to action:

"We urge that the European Commission promote, and the member-states adopt, new policies to foster the development of advanced-degree programmes at our major universities for this emerging field of data science. We also urge the member-states to include data management and governance considerations in the curricula of their secondary schools, as part of the IT familiarisation programmes that are becoming common in European education." (European Commission, 2010, p. 32)

Additional studies and efforts by Library and Information Science schools, as highlighted by the "Preparing the workforce for digital curation: The iSchool perspective" panelists at the 9th International Digital Curation Conference, also continue to help discovering, improving, and expanding the education and development opportunities for information professionals.

Among the many different programs that are implemented to address the need to prepare information professionals, the Graduate School of Library and Information Science (GSLIS) at the University of Illinois at Urbana-Champaign (UIUC) takes a "leading role in both data curation education and research" (GSLIS The iSchool at Illinois, n.d., p. 2) by offering a degree specialization in Data Curation. Specifically, the Specialization in Data Curation program was developed in 2006 through a grant from the Institute of Museum and Library Services (IMLS), and can be earned "either as part of an ALA-accredited Master of Science (MS) degree or, for students who have already completed their master's degree, as part of a Certificate of Advanced Study (CAS) degree" (GSLIS The iSchool at Illinois, n.d., para 3). The program defines Data Curation as the "active and ongoing management of data throughout its entire lifecycle of interest and usefulness to scholarship, science, and education" (Cragin et al., 2007). By focusing on the collection, representation, management, and preservation of data, the program "brings together library and information science and archival theory as well as digital technologies for information discovery, access, and re-use" (GSLIS The iSchool at Illinois, n.d., p. 2). In addition, since UIUC's Center for Informatics Research in Science and Scholarship (CIRSS) oversees the Specialization in Data Curation program, the Center injects further data curation education and research opportunities into the program. As a result, UIUC GSLIS expects that graduates of the Specialization in Data Curation program will be trained and develop the necessary expertise to be employed "across a range of institutions, including museums, data centers, libraries and institutional repositories, archives, and throughout private industry" (GSLIS The iSchool at Illinois, n.d., para 3) to help curate and manage the data and the associated requirements.

While the knowledge and skills gained from the Specialization in Data Curation program can be applied to all data relating to sciences, humanities, and social sciences, students can select elective courses beyond the three required core courses to personalize the program experience to their personal interests and professional goals. In addition, students can participate in projects hosted by CIRSS to gain hands-on curation experiences. An example project is the Data Curation Education in Research Centers (DCERC) program. The project is led by UIUC in collaboration with the University of Tennessee and the National Center for Atmospheric Research (NCAR), a premier national research center with state-of-the-art data operations and services. The goal of the DCERC program is to develop a sustainable and transferable model of data curation education for masters and doctoral students in Library and Information Science. To achieve this goal, a key part of the program has been summer internships for masters students at NCAR. During the internships, the students are expected to complete a data curation project relating to scientific research at NCAR. In addition, the students are paired with both science and data mentors based on their areas of interest and technical skill level in order to further understand NCAR's research and data environment. Furthermore, the internship offers the students other opportunities to participate in activities, such as conferences, talks, and workshops, which allow the students to explore additional topics relating to the practices and policies of data curation and management.

For the author, her experience at NCAR through the DCERC program allowed her to work with the mentors at the NCAR Computational and Information Systems Laboratory's (CISL) Research Data Archive (RDA) and NCAR Research Applications Lab (RAL) in order to make a unique climate reanalysis dataset publicly available, accessible, and usable. The dataset contains three-dimensional hourly analyses in netCDF format for the global atmospheric state from 1985 to 2005 (a total of 184,080 files) on a 40km horizontal grid (0.4°grid increment) with 28 vertical levels. As a result, the dataset provides detailed representation of local forcing and the diurnal variation of processes in the planetary boundary layer to allow and promote studies of new climate characteristics. During the project, the author focused on three specific areas of the data curation process: data quality verification, metadata descriptions harvesting, and provenance information documentation. When the curation project started it had been five years since the data files were generated. Also, although the Principal Investigator (PI) had generated a user document, the document had not been maintained. Furthermore, the PI had moved to a new institution, and the remaining team members were reassigned to other projects. These factors made data curation in the author's focus areas especially challenging. As a result, the project provided the author a realistic environment to understand and practice the methodologies for resolving data curation issues in a scientific research setting. Overall, the author was able to make the dataset available, accessible, and usable through the data's landing page at RDA at the end of the eight-week internship. The project illustrated that it was essential for the proper and dedicated resources to be invested in the curation process in order to give datasets the best chance to fulfill their potential to support scientific discovery. Equally important, the project team also reflected the following key experiences with the data curation process:

- Data curator's skill and knowledge helped make decisions, such as file format and structure and workflow documentation, that had significant, positive impact on the ease of the dataset's management and long-term preservation.
- Use of data curation tools, such as the Data Curation Profiles Toolkit's guidelines, revealed important information for promoting the data's usability and enhancing preservation planning.
- Involving data curators during each stage of the data curation life cycle instead of at the end could improve the curation process' efficiency.

As the data and their associated management and curation requirements grow, the expertise of the trained information professionals will also become more important to meet the challenges. It will be important to continue to raise awareness and emphasize the need to implement the framework and policies to support the training and professional development of information professionals in the area of data curation. Meanwhile, through her dedicated academic coursework in data curation at UIUC and practical experiences with the Research Data Archive at NCAR, the author provides an example of how the current effort is already making positive impact on the next generation of information professionals who are willing and welcome the opportunities to take on the responsibilities.

References

Cragin, M.H., Heidorn, P.B., Palmer, C.L., & Smith, L.C. (2007). An Educational Program on Data Curation. Poster, Science and Technology Section of the annual American Library Association conference. Washington, D.C., June 25, 2007. Available: http://hdl.handle.net/2142/3493.

European Commission. (2010, October 6). Riding the wave – How Europe can gain from the rising tide of scientific data – Final report of the High Level Expert Group on Scientific Data. Retrieved from http://ec.europa.eu/information_society/newsroom/cf/dae/document.cfm?action=display&doc_id=707

Graduate School of Library and Information Science The iSchool at Illinois. (N.D). Specialization in data curation. Retrieved from: http://www.lis.illinois.edu/academics/degrees/specializations/data_curation

Graduate School of Library and Information Science The iSchool at Illinois. (N.D.). Specialization in data curation program overview. Retrieved from http://webdocs.lis.illinois.edu/comm/recruitment_pdfs/Specialization-Data-Curation.pdf

Hedstrom, M., Larsen, R., Palmer, C., DeRoure, D., & Lyon, L. (2014, February 25). Preparing the workforce for digital curation: The iSchool perspective. Panel discussion presented at the 9th International Digital Curation Conference, San Francisco, CA.

Research Information Network and the British Library. (2009, November 2). Patterns of information use and exchange: case studies of researchers in the life sciences. Retrieved from http://www.rin.ac.uk/our-work/using-and-accessing-information-resources/patterns-information-use-and-exchange-case-studie

Commentary

A useable simulation model archive: Does it really exist?

edit

Mark E. Harmon, Richardson Chair and Professor, Department of Ecosystems and Society, Oregon State University, Corvallis, OR, 97331
Edward B. Rastetter, Senior Scientist, The Ecosystems Center, Marine Biological Laboratory, Woods Hole, MA, 02543

At some point in a career everyone no doubt wonders what their legacy will be. That is certainly the case for the two of us, as we are nearer the end than the start of our careers as ecologists. While this might seem a narcissistic obsession, there is a practical side as well for everyone else. One of the hallmarks of humans as a species is the ability to pass knowledge, skills, technology, and experience from one generation to the next. So it is sensible to wonder what will be passed along from one scientific generation to the next. For scientists, the hope is to pass along critical concepts, facts, technologies, samples, and observations. If this does not happen then each generation must start over and scientific progress will be delayed because old findings have to be rediscovered.

It is not enough to want to pass a legacy along - there must be mechanisms to enable this process. In thinking over the various parts of our own legacies, we are confident that for some parts the mechanisms are well established. Most anything appearing in a journal article or book, be it either data, concepts, or knowledge, will be archived in various libraries, online databases, and other well established and redundant systems such as JSTOR, Questia, SCOPUS, the Web of Science, and even the Library of Congress. Ecological data itself can be entered, documented, and retrieved using several systems such as CDIAC (Carbon Dioxide Information Analysis Center), the US LTER's NIS (Network Information System) and the Ecological Society of America's Ecological Archives, all of which will continue to evolve but at least are in place. The ability to pass along physical samples is less sure, however. Systems to archive, curate, and retrieve this material are widespread, ranging from classical organism collections in museums to site-based systems such as Hubbard Brook's Sample Archive. The main issue is whether or not one can take advantage of these systems based on one's location and type of sample. What we do not seem to be able to find is a system to enable legacies for ecological simulation models (Thornton et al 2005). The rest of this commentary describes what we think such a system would look like and the steps required to create it.

Before describing this system, let us first be clear why we think simulation models are in particular need of attention. There are many forms of models used in ecology. Analog models (e.g., microcosms) are generally not saved, although no doubt some are in museums and many are described in publications. Since being replaced by digital models there is probably little need to be concerned about the ability to recreate analog models outside of an educational setting. Conceptual (e.g., N-saturation model of Aber et al. 1989) and analytical (e.g., NEE model of Shaver et al 2007) models are usually relatively simple and well documented in publications. The same is usually true for empirical models as the main purpose of many publications is to use data to develop these relationships. While the data to develop empirical models is often inadequately presented in publications, there are at least systems (as noted above) to store everything from the raw to the processed, cleaned data. Simulation models are generally more complicated than analytical ones, and although described in publications, there is generally not enough room to do this fully; hence much of the information needed to use or recreate them exists outside the publication system. Moreover, unlike either an analog or analytical model, digital simulation models are generally not simple to recreate and given the limited description in publications, may be impossible to recreate exactly from that source alone. This is unfortunate because simulation models have become a way to synthesize ecological knowledge, explore integrative hypotheses, and analyze complex systems. As these reflect relatively new ways to think about ecological systems, failing to pass simulation models from one generation to the next is potentially an extremely unfortunate situation that could slow progress in ecological sciences. In a sense it would be similar to every generation having to reinvent the elemental analyzer or some other critical piece of technology that we currently take for granted.

Developing a system to usefully document, archive, and retrieve ecological simulation models will involve considerable thought. Part of the complexity of this effort is reflected by the fact simulation models are really an amalgamation of concepts, hypotheses, data, and technology. Fortunately parts of other systems can be reused and modified to create this new model archiving system. For example, data are usually used to drive simulation models and data is a primary model output. Documentation of these model-related data can take advantage of existing systems such as LTER's NIS that document, archive, and retrieve data (spatial or non-spatial). Model parameters can also be described using these systems; however, there is additional information on the source, transformation (often parameters are derived from data and this process needs to be described), uncertainty, and other aspects of model parameters that need to be added. It is extremely useful to understand how sensitive a model is to changes in driving variables and parameters. While some of this information may be described in publications, detailed examinations of sensitivity often undertaken by model developers are generally not formally documented. It would be useful for future users if a sensitivity analysis was part of every simulation model's documentation (e.g., Grimm et al. 2014). As simulation models are developed it is not unusual to have multiple versions of models and while it may not be practical to save every version, those that represent significant milestones (e.g., either a major change in functionality or publication of a key analysis) of development should be archived. Fortunately conventions developed for other forms of software development can be used. While storing of the computer code (i.e., source and executable files) will not be challenging per se, one cannot expect computer code created on one operating system environment to automatically be useable on another or for the code to run under some future operating system. Therefore, in addition to the code itself, it may be necessary to archive the operating system in which that code was developed, which in turn might mean also physically saving the hardware able to run that operating system. Alternatively, new technology now allows the development of virtual machines to simulate one operating system and associated hardware on another. These virtual machines can be more easily passed from one generation of system to the next. Finally, as the point of archiving simulation models would be to use them again, this process would be greatly helped by archiving input and output data that can be used to test if the recreated model is acting as expected and to serve as a template for formatting new parameter and driver files to be used with the archived model.

We envision an archival system where not only the full information needed to recreate the model is available, but the model itself is available and usable under any future operating system. Imagine, for example, being able to rerun the original Botkin et al. (1972) JABOWA simulations for Hubbard Brook from your computer without having to recode the model. Bear in mind the original JABOWA was developed on punch cards for an IBM mainframe computer, a system not currently available to anyone outside of a museum. In the archive we envision, the JABOWA code and a simulated IBM operating system would be archived on a virtual machine so that the model could be rerun not only with the original input files but with any newly created input file in the same format.

We have few illusions as to the challenges to be faced in developing the proposed system and while issue has been previously noted by others (Thornton et al 2005), little appears to have been done to address it. This indicates to us that perhaps one of the largest challenges is to have scientists and funding agencies recognize the need for such a system and to understand that it would be different from what currently exists. We may be mistaken, but in our conversations with others we have the impression that there is a widespread belief that such a system already exists (how could it not?) or that current systems for data will be sufficient. We are not really sure a general system exists and suspect those that might are not sufficient without some modification. Another challenge is that a system to document, archive, and retrieve simulation models will cost time and resources in development and in use. Those using data systems will understand that proper documentation and archiving of data can add 25-35% effort to a project, which in a fixed budget world means fewer publications and presentations. We would expect the same costs for a simulation model system; unless scientists and funding agencies support these costs in terms of lower short-term productivity there will be reluctance of simulation model developers to bear them. This would be indeed unfortunate in that failure to accept these short-term costs will likely come at the expense of long-term productivity.

References

Aber, JD, KJ Nadelhoffer, P Steudler, and JM Melillo. 1989. Nitrogen saturation in northern forest ecosystems. BioScience 39: 378-386.

Botkin, DB, JF Janak, and JR Wallis. 1972. Source Some Ecological Consequences of a Computer Model of Forest Growth. Journal of Ecology 60:849-873.

Grimm, V, J Augusiak, A Focks, BM Frank, F Gabsi, ASA Johnston, C Liu, BT Martin, M Meli,V Radchuk, P Thorbek, and SF Railsback. 2014. Towards better modelling and decision support: Documenting model development, testing, and analysis using TRACE. Ecological Modelling 280:129-139.

Shaver, GR, LE Street, EB Rastetter, MT van Wijk, and M Williams. 2007. Functional convergence in regulation of net CO2 flux in heterogeneous tundra landscapes in Alaska and Sweden. Journal of Ecology 95: 80--817.

Thornton, P. E., Cook, R. B., Braswell, B. H., Law, B. E., Shugart, H. H., Rhyne, B. T., and Hook, L. A. 2005. Archiving numerical models of biogeochemical dynamics. Eos, Transactions AmericanGeophysical Union 86: 431-431.

Good Tools And Programs

Increasing Readability of R Output with Markdown

edit

--John Porter, Virginia Coast Reserve LTER

The R statistical package has many statistical and graphical capabilities and a large and active user base, making help easy to find on the Internet. However, raw R output is relatively bare-bones making it difficult to display in easily interpretable forms – especially after a long time has elapsed.

One solution that helps increase the readability of R is Markdown. Using Markdown you can create dynamic documents that integrate text, R code, R results and R graphics into a single HTML, PDF or Word document. Markdown is integrated into the popular RStudio user interface for R, but can also be run directly from the command line.

Describing Markdown may best be shown with a simple example. Here is the "default" R MarkDown (.rmd) document supplied by RStudio:

````
---
title: "R_Markdown_Demo"
date: "Thursday, December 04, 2014"
output: html_document
---

This is an R Markdown document. Markdown is a simple formatting syntax for
authoring HTML, PDF, and MS Word documents. For more details on using R
Markdown see <http://rmarkdown.rstudio.com>.

When you click the **Knit** button a document will be generated that
includes both content as well as the output of any embedded R code chunks
within the document. You can embed an R code chunk like this:

```{r}
summary(cars)
```

You can also embed plots, for example:

```{r, echo=FALSE}
plot(cars)
```

Note that the `echo = FALSE` parameter was added to the code chunk to
prevent printing of the R code that generated the plot.
````

Hitting the "Knit HTML" button in RStudio or using the rmarkdown::render function from the command line, produces an HTML output that looks like this:

R_Markdown_Demo

Thursday, December 04, 2014

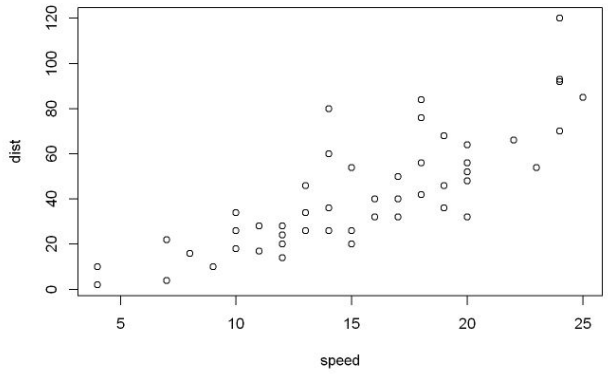
This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
summary(cars)
```

```
##      speed           dist
##  Min.   : 4.0   Min.   :  2
##  1st Qu.:12.0   1st Qu.: 26
##  Median :15.0   Median : 36
##  Mean   :15.4   Mean   : 43
##  3rd Qu.:19.0   3rd Qu.: 56
##  Max.   :25.0   Max.   :120
```

You can also embed plots, for example:



Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.

The Markdown language is relatively simple. R code appears in triple single backquotes labeled with {R} and options for whether the code will be echoed or not. A single asterix (*) can be used to produce *italics* (e.g., *italics*) double asterix "**" delimit the start and end of bold - thus **bold** comes out **bold**. Although not show in the example, proceeding text with a # makes it into a large heading (## and ### make successively smaller headings) and preceeding text with a single hyphen ("-") makes it appear as a bulleted list item and a plus ("+") a numbered list item. There are also options for including tables, hyperlinks and even LaTeX style equations. See http://rmarkdown.rstudio.com/authoring_basics.html for details.

Markdown makes it relatively easy to generate ready-to use outputs showing data displays for posting on the web or including in reports as WORD documents. However, I also use it to provide more attractive and readable commentary regarding my R code than the simple comments embedded in the code itself. If using RStudio, you can also select and run individual chunks of R code for testing and debugging. This is helpful because "kniting" large and complex documents takes slightly longer than running the R code alone.

Capturing Location Data for Field Plots

edit

John Porter (VCR)

Recently I sent out a query to LTER Information Managers requesting information about good ways to collect field research location data from users. For point data, this is relatively easy, since only a single pair of coordinates is required and these can be gotten from any Global Positioning System (GPS) unit or even smart phone,or obtained by noting coordinates displayed in mapping tools such as Google Earth. However, things rapidly get more complicated when complex polygons of plot boundaries are required for multiple plots and where there are many different data providers. My goal is to create (or find) an easy-to-use system that will allow a diverse array of investigators to generate points, lines or polygons for study sites that can submitted for import into GIS software or a database for the purpose of populating metadata and research site maps.

I received several helpful responses. Fox Peterson at the H.J. Andrews was first. She proposed a system based on using the outline and edge-detection functions in Matlab to convert outlines or areas drawn using a “paintbrush” on top of a high-resolution, georeferenced aerial image of the site drawn. She cited information on the Matlab Mapping Toolkit (http://www.mathworks.com/help/map/_f7-12036.html) and additional information on edge detection (http://www.mathworks.com/discovery/edge-detection.html).

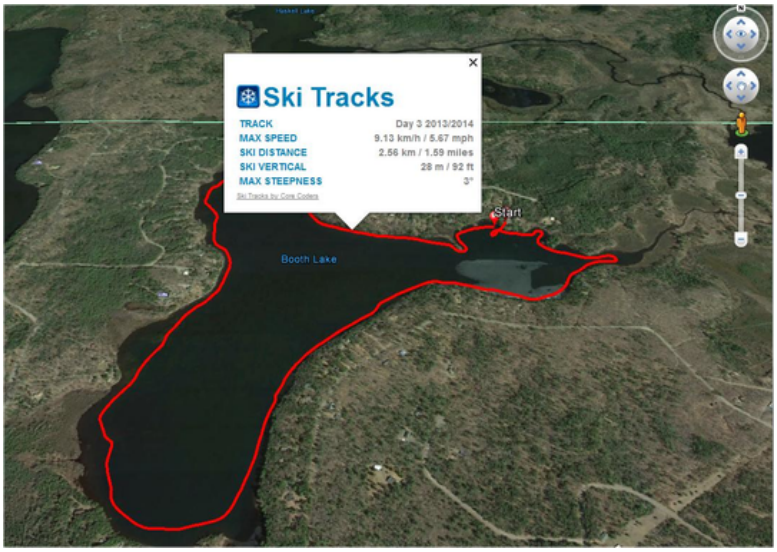
Gastil from the Moorea Coral Reef LTER recommended looking at
SeaSketch – a collaborative platform for GeoDesign (http://mcclintock.msi.ucsb.edu/projects/seasketch). This developing tool goes well beyond capturing boundary coordinates, providing automated feedback on potential impacts of manipulations and forums for sharing and discussing sketches. The tool is aimed at marine environments, but she suggested that the developers might be open to wider collaboration.

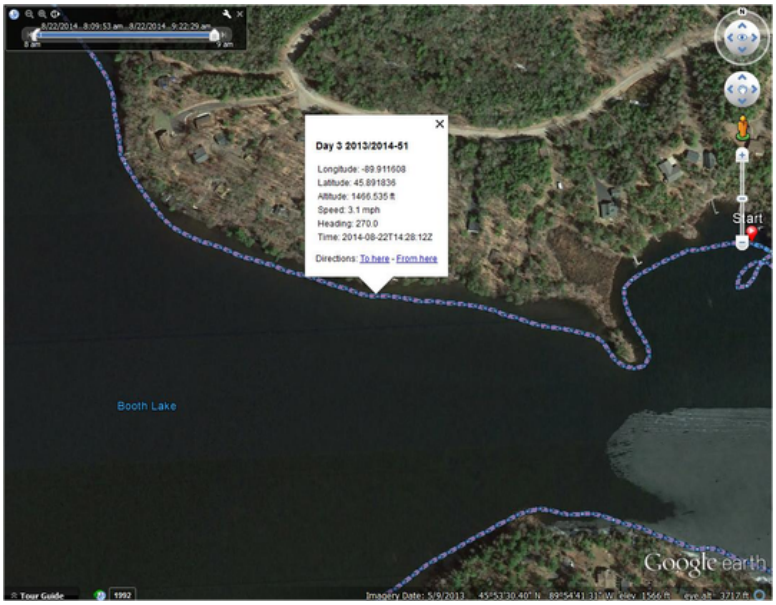
Theresa Valentine from the H.J. Andrews LTER provided some useful tips on dealing with capturing the individual sampling points associated with gridded plots. She uses a GPS location at one corner of the grid, along with sample point spacing, number of plots and grid orientation to automate generation of individual sampling points using the “fishnet” tool in ArcGIS. For features that are clearly visible from aerial imagery, digitization off of high-resolution photos may be used, or field surveys using optical surveying tools may be converted from a coordinate geometry to conventional coordinates. For irregular linear features, such as trails, they combine GPS data from multiple trips down the trail. In a forested environment, dropout of GPS signals is a recurrent problem, so collection of key points such as switchbacks or turns may require revisits.

She further discussed the desirability of high-resolution GPS data so as to match up with high-quality LiDAR-based digital elevation data. Raising the antenna well up into, or over, the canopy helps in a forested environment. Additionally, picking dates for surveys with good satellite geometries using planning tools and differential correction to a local base station help achieve the best accuracy. They do a GPS foray every summer to collect accurate GPS coordinates of study site locations.

Additionally, John Vande Castle from the LTER Network Office offered a wealth of information regarding the use of Global Positioning System data and tradeoffs between the KML data format used by Google Earth and the GPX exchange format for GPS data. In addition to tracks recorded directly from dedicated Garmin and other

GPS devices, he uses apps, such as “Ski Tracks” or Google’s “My Tracks,” that keep a record of GPS points (a track). The apps generate KML files that contain information specific to the app along with the coordinates. For example an app aimed at skiing gives information on speed and slope (Figure 1). He noted the benefit of transforming the KML into GPX in order to obtain the coordinates and line distances in a more consistent, less esoteric, format for general use (Figure 2). He then can manipulate the data with the (now free) Garmin BaseCamp software (Figure 3) or with other GIS tools.





Additionally, I investigated using Google My Maps (https://www.google.com/mymaps), ScribbleMaps.com and Google Earth as potential tools. Each of these had some advantages and some drawbacks. ScribbleMaps shares many similarities with a large number of map editing sites built on top of Google or ESRI frameworks such as: Mapbox (https://www.mapbox.com – allows GPS file import), ZeeMaps.com (points only), Waze.com, National Geographic MapMaker

Interactive (http://mapmaker.education.nationalgeographic.com/, no way to save polygons), Map Maker (http://mapmaker.donkeymagic.co.uk/, points only) and many others.

Feature	Google "My Maps"	ScribbleMaps	Google Earth
Accounts	Required	Optional, but all anonymously saved maps are public	None, but software must be installed
Editor	Easy to use, all types of features can be labeled through editing	Easy to use, but only points may be labeled with a title.	Moderately easy to use. Editing is a secondary function so controls are harder to find.
Export Capabilities	Exports to KML retaining names and other annotations	Can export XML in either KML or GPX forms. Polygons and lines have unique IDs but these are set by the system and hidden from the user.	KML/KMZ (zipped KML) via "Save Place as"

All of the online tools use available satellite and aerial photo services, which vary in their accuracy from location to location (fortunately for the VCR/LTER there is a high agreement between GPS and image-based sources). However, before using the online tools, it is a good idea to cross-check the accuracy for known points in your area. The most common export format, KML, is an XML format that can be transformed to extract location information for use in databases and metadata. However, the "richness" of the KML varies among services. For example, ScribbleMaps allows creation of points, lines or polygons, but user-supplied identification information is only exported for points. In contrast, Google Earth exports name, description and style information.

I would like to say that I identified the "perfect" system. All of the suggestions have merit and there is clearly rapid evolution occurring in the online map making community. However there are still gaps that require manual intervention (from collection of GPS data to data conversions). Nonetheless, there are a number of tools and approaches that could be applied to the field data geolocation problem.

Good Reads

Visualization Blog "Cool Infographics"

edit

Theresa Valentine (AND)

I stumbled on this interesting company called Cool Inforgraphics, The site uses charts and graphs to communicate data and turn data into information. Rany Krum, President also has written a book called Cool Inforgaphics: Effective Communication with Data Visualization and Design. All the figures in the book are available at http://www.coolinfographics.com/figures/ . Check out the Awesome Tower of Beer, and The Lifespan of Storage Media, What's so hard about Search?. Chapter 7 has a list of design resource links, including Inkscape, Chartle, DIY Chart (Do it yourself), and lots of others. Many are open source options. One last link is A Periodic table of Visualization Methods http://www.visual-literacy.org/periodic_table/periodic_table.html where you will find interesting methods such as tree map, force field diagram, argument slide, and the heaven n hell chart.