LTER Databits

Information Management Newsletter of the Long Term Ecological Research Network

Fall 2009

The 'Fall 2009 Databits' publication is here and is full of great articles for your enjoyment. In this issue you will find a 'Feature Article' describing some recent revelations from a cross-site exchange among LTER information managers as well as range of 'News Bits' articles that report the progress from four key network information management working groups, reveal various site methods for handling GIS metadata, and document one sites' fallout following a hurricane. There are also 'Good Tools and Programs' articles detailing an open source plotting library program and Firefox extensions for web developers. The 'Good Reads' section showcases a number of interesting publications that may interest you. And finally, this issue has a 'Commentary' outlining one potential continuing education opportunity for information managers. Enjoy!

Featured Articles

ClimDB/HydroDB (ClimHy) Database Migration to LNO
LTER IMC and IMExec: 2009 Progress and Planning
Experiences from an Information Management Cross-Site Visit

Commentary

Continuing education options for Information Managers

News Bits

GIS Working Group Report
IMC Governance Working Group
LTER Unit Registry: Products and Processes
Developing a Controlled Vocabulary for LTER Data
Beware the Ida's of November
How do you put your GIS metadata in EML format?

Good Tools And Programs

FireBug: Web Customizing To Fit Your Needs
Matplotlib: An Open Source Python 2-D Plotting Library

Good Reads

A book for data people
Nature special issue on Data Sharing
Standards and Data Sharing in Ecology
Identifying Best Practice and Skill for Workforce Development in Data Curation
The Ghost Map by Steven Johnson

Calendar

Events

Featured Articles

ClimDB/HydroDB (ClimHy) Database Migration to LNO

edit

- Suzanne Remillard (AND) and Don Henshaw (AND)

ClimDB/HydroDB is on the move. ClimDB/HydroDB (ClimHy) is a web harvester and data warehouse that provides uniform access to common daily streamflow and meteorological data through a single portal (Henshaw et al. 2006). ClimHy, which has been hosted by the Andrews LTER since 2003, is being migrated to the LTER Network Office (LNO). This migration is occurring to relieve the support and administrative burden from the Andrews LTER site and LNO welcomes hosting this key Network Information System (NIS) module. LNO hosting will improve efficiency in the eventual integration of this module into the new NIS architecture.

The timeline of migration is to have all sites using the LNO application by early 2010.The database administration will still be performed by the ClimHy administrator, Suzanne Remillard (AND), until the LNO hires the new Network Information Manager (NIM). The LNO will take full responsibility for management and curation by Spring 2010. The Andrews IM will continue to be available to provide consultation until the NIM is fully trained. The Andrews ClimHy system includes the database server, file server and web server. As of now, an identical system is established at the LNO. Significant testing has been conducted to ensure stability of the system at the LNO.

What does this mean to participating sites?

Most of the migration will be seamless to participating sites and public users, however please note the following. The Andrews ClimHy database will remain the legitimate production database for harvest, access, and metadata entry until the exact date for migrating the production database (date will be determined and announced soon; probably early February 2010). At that announced date, the LNO server will become the production ClimHy server and the Andrews ClimHy server will be disabled. In the meantime, sites can begin testing both manual and programmatic harvesting to the LNO ClimHy server.

The new LNO ClimHy website is currently operational and participating sites can initiate test data harvests to the LNO server. The following URLs are active and ready for testing:

New webpage access to ClimDB/HydroDB site at LNO:

Participant Page: http://climhy.lternet.edu/harvest/harvest.htm

Public Data Access Page: http://climhy.lternet.edu

The migration will affect participating sites that run dynamic scripts to harvest. These sites can immediately begin testing the new LNO harvest URL in their scripts. It is suggested that scripts be revised as soon as possible to allow simultaneous harvest to both servers, at least until the production server is moved from the Andrews to LNO.

The new LNO harvest URL is: http://climhy.lternet.edu/harvest.pl/harvest.pl?module=<#>&site=

where,

<#> = URL option number (1 or 2)

= Three letter LTER Site code (i.e., AND for Andrews)

The 'last harvest' web service allows participants to provide site, station, and parameter to retrieve the last harvested date (contact the ClimHY administrator for more information). This web service has also been migrated and can be accessed at LNO: http://climhy.lternet.edu/wambam/services/climdb_raw_handler.pl Sites should keep in mind that all access and harvests should still be directed to the Andrews production server at this time, and that use of the LNO Server is only for testing until the actual production server change date Suzanne Remillard (AND) will work with LNO on scheduling auto-harvests from the LNO server. Wade Sheldon (GCE) has modified his system to harvest USGS sites to both the Andrews and LNO servers. Participating sites have probably noticed ClimHy reports coming from both servers. Stay tuned for future announcements regarding this migration.

Reference

Henshaw, Donald L.; Sheldon, Wade M.; Remillard, Suzanne M.; Kotwica, Kyle. 2006. CLIMDB/HYDRODB: a web harvester and data warehouse approach to building a cross-site climate and hydrology database. In: Proceedings of the 7th International Conference on Hydroscience and Engineering (ICHE-2006); Philadelphia, PA. Philadelphia, PA: Drexel University, College of Engineering: [Online]. Available: http://hdl.handle.net/1860/1434

LTER IMC and IMExec: 2009 Progress and Planning

edit

Margaret O'Brien (SBC) and Don Henshaw (AND)

The Information Management Committee (IMC) meeting at Estes Park was attended by over 50 site information managers and guests, and as usual, was the event at which we mark progress. The IMC continued its dialog on its own governance, as well as its interactions in network and community processes and project development models which suit both IMC and network CI needs. Our product-oriented working groups have been very productive over the past year, yielding results that both further integrate information management into network science, and enhance site functionality. The "Units Dictionary" group plans to launch its web services in 2010, allowing sites either to retrieve measurement units or to implement the dictionary and registry locally. Parts of our "Controlled Vocabulary" are already in place for dynamic searches in the network data catalog, and activities are now centered on the relationship of an LTER vocabulary to other efforts concerned with keyword and vocabulary modeling. The GIS working group has developed recommendations for spatial data to be reviewed by the IMC, and was given a post-ASM meeting award to complete their Google maps application, "LTERmaps." See the contributions from those groups elsewhere in this issue.

Two new working groups were also established at the ASM, both of which highlight site-functionality and network partnerships. The "EML Metrics" working group is a community-wide collaboration of information managers and ecoinformatics programmers spearheaded by the LTER. Its goal is to create metrics and reporting tools for EML datasets to assure integrity and quality for automated ingestion by systems such as the NIS. A second group, "Network Web Services," will focus on the functionality of LNO databases and the exchange of content between central databases and sites' IM systems - functionality which is a key part of the LNO operational plan. The web services group will make recommendations which directly benefit site activities, including population of local web sites from network databases, providing standardized content for EML metadata, and streamlined contributions from sites to the network.

Since the September ASM, IMExec has been busy conducting bi-monthly meetings via videoconferencing (VTC). These meetings have largely been devoted to planning IMC activities such as the monthly virtual water cooler sessions, considering future production workshop and training workshop possibilities, coordination of site activities, and most recently, reviewing the draft LTER Network Office (LNO) Operational Plan. Once accepted, the Operational Plan will govern activities at LNO for the next 6 years, including development of the Network Information System (NIS), and is highly relevant to site information management efforts.

IMExec has considered the coordination of cross-site activities should additional funding be provided to sites. Funding could be directed toward improving the quality and availability of EML metadata and data or improving network-wide standardization to facilitate increasing use of site data in synthesis projects. IMExec has outlined activities that might better enable dynamic loading of site data into Network architecture or promote standardization of data and conformance to best practices including:

- Improvements in completeness and accuracy of metadata
- Evaluation and improvement of linkages between metadata and data
- Evaluation and improvement of data quality
- Automating EML generation for new data types (e.g. GIS, non-tabular, or non-LTER data)
- Development of software tools or stylesheets to improve the usefulness of EML to LTER sites and end-users
- Adoption of emerging Network standards (e.g. keyword, attribute, or taxonomy vocabularies)
- Initiatives promoting or improving standardization and integration with LNO developments (e.g. refinement or expansion of current best practices for EML)
- Exploring dataset annotation with emerging semantic tools

There may be an advantage to co-development of approaches among sites and it is likely that workshops will be necessary to develop standardized approaches and best practices. The NIS Advisory Committee (NISAC) has proposed that taskforces of mostly LTER personnel be assembled to work out required standards and best practices, and efforts might be combined with IM production workshops or other synthesis working groups. IMExec will be discussing potential workshops to best prepare sites for participation in the NIS at the annual IMExec meeting in February. With this in mind we will be exploring necessary metrics for evaluating site EML and consider more advanced tools for validation of site data and metadata. We will also be examining the NIS framework and development timeline as outlined in the Operational Plan to help guide our future planning. This is a challenging, busy and exciting time for all site information managers and

the IMExec Committee. The IMC co-chairs have been successful in sharing the workload and stay in close communication, and we look forward to upcoming workshops and new opportunities for improving and better integrating our site and network information systems.

Experiences from an Information Management Cross-Site Visit

edit

Nicole Kaplan (SGS) and Karen Baker (PAL/CCE)

At our Information Management annual meetings we traditionally schedule time to speak with other information managers about issues, challenges, and practices. We find time to discuss troubles, workarounds, and change, but the time available for informal communication is never enough. A recent cross-site exchange between SGS and CCE/PAL provided an additional window of time that allowed each of us to examine more closely how information management is performed at another site. Benefits of such a visit include an opportunity to hear about design and implementation details from another LTER site, to gain insight from comparative study of site differences and similarities, and to explore unexpected topics of joint interest.

Initially, the CCE/PAL Ocean Informatics group was able to suggest the visit because funds for a cross-site exchange were included in a supplement request. SGS, however, offered to support a large part of the visit from site funds. Ocean Informatics (OI) is a conceptual umbrella for the work of two LTER sites, PAL and CCE. The OI team's focus on 'Continuing Learning' was evident as one member shared her plan for telecommuting in order to complete a Masters Degree (see article by LYarmey). The whole group gathered for a weekly discussion group with attendees including developers, technicians, analysts and a science studies researcher. In addition, several times, as subsets of the team, we gathered at the Ocean Informatics design table. A morning spent demoing our respective web sites and information systems prompted a number of discussions. This served as valuable peer-review process as functions and designs have been implemented within all our systems to discover data and information that incorporate similar aspects of community supported standards, such as EML and LTER controlled vocabularies, but were built on very different information architectures.

For the site-to-site exchange, we initially took time to review how data are collected, structured, stored, managed, and delivered from differing physical ecosystem and are handled within differing local-level information ecosystems. This was a very worthwhile experience as strategies for managing different data types using various data models emerged as we shared approaches and tools used at each site. We shared a wide-ranging set of specific elements relating to site information management by:

- Reviewing IM sections of proposals for SGS, PAL and CCE
- Sharing metaphors and analogies of how we work in our teams and within the network
- Discussing future plans for managing new data types and increased data flows

With two sites with distinct development trajectories, each could use input from the other to prompt definition and articulation of our own particular set of site information management arrangements. In addition, these discussions assisting in planning future goals for site information management systems and expertise to support proposed science within the upcoming LTER renewal proposal submissions. Finally, a site's support of a cross-site collaboration with colleagues is viewed by the local PIs as an investment in professional development, an opportunity for information and technology transfer, and participation at the network level. We have begun to consider a variety of formulations that might replicate the beneficial aspects of a site-to-site exchange. One possibility, would be to create a one-time Peer Review Panel. Such a panel would be an opportunity to gather together several professionals to strategically explore and critically assess IM. We found time to discuss network-level activities including the challenge of integrating diverse data that is not interoperable despite falling into the same category or theme such as net primary productivity. At a network level, we shared topics by:

- Articulating the changing requirements associated with information management
- Considering professional meetings for information managers
- Exchanging experiences in submitting data to Ecotrends
- Discussing plans for enacting the unit registry at local and network level

In our case, there was an added benefit in that the visit was planned to precede the All Scientist Meeting so that, as co-members of the Governance Working Group, we were able to develop materials, discuss best practices, and plan agenda items during the exchange. We conferred on a figure central to discussion of site-network governance at this year's IMC. Time was also spent on design of an LTER Information Management History database (HistoryDB) and creation of a poster on the same topic for the ASM.

Working at geographically distant and organizationally distinct sites meant it was no surprise to find that we approached problems using different expertise and had tasks arranged with different divisions of labor. A site exchange facilitates learning and professional growth, mentoring between seasoned and newer team members or students, and dialogue on approaches to project management that ensure efficient and effective data practices. Our particular work plan included tasks relating to local site arrangements, Information Management Committee activities, and network level issues. The visit was rounded out by visiting the Ocean Informatics Computational Infrastructure facility with its servers and surfboards, visiting the Birch Aquarium with its fish tanks and education exhibits, and participating in the #1 UC Yoga Club.

After working on information management for an LTER for more than 7 years (Nicole Kaplan) and more than 18 years (Karen Baker), the visit had the feel of what we called a 'mini-sabbatical'. It proved both a productive and a fun way to learn, assess, design and plan collaboratively. Two noticeable aspects of the cross-site exchange were the ability to step back to consider local work within a broader context and to become re-energized by new ideas and by re-imagining the role of information management. We would urge other information managers to consider planning and requesting support from your site(s) for a cross-site visit as an opportunity to share expertise and perspectives, discuss similar challenges and develop potential solutions, address how network level efforts or concerns affect different sites, and build partnerships with like-minded professionals. If you are considering requesting support for a cross-site visit or incorporating one into a supplemental proposal, be sure to state your goals and intentions clearly and share your accomplishments with your site, as well as the information management community. Our sites and our network can benefit from these experiences.

Commentary

Continuing education options for Information Managers

edit

Lynn Yarmey (CCE/PAL)

I am currently in the midst of completing the first semester of my Library and Information Science (LIS) Master's program with a concentration in Data Curation through the iSchool (http://www.ischools.org) at the University of Illinois, Urbana Champaign (http://www.lis.illinois.edu). I am a 'distance learner' so all of my classes are conducted online, each class meets once per week using the Elluminate software package. The experience, while couched in the library context and perspective, has offered a great deal in terms of insights, theory, literature and skills that directly pertain to my work with the LTER and at Scripps Inst. of Oceanography. After applying, I was awarded a Data Curation Education Program fellowship that enables me to participate as a full-time student; my classes so far have included two core library classes as well as Foundations of Data Curation and Systems Analysis and Management. Topics in these first core classes have covered the breath of Information Science, from cataloging, subject access and indexing on the library side, to academic and science-related digital data

lifecycles, scholarly communication, data collections, archiving and preservation, ontologies, metadata and provenance. Next semester I am looking forward to taking a class in databases and one about building instruction systems. A third class on collection development will be taught by Dorothea Salo, an institutional repository librarian whose work I admire. Guest lecturers in my classes this semester have included Ruth Duerr from the National Snow and Ice Data Center (http://nsidc.org/) and Gail Steinhart representing the DataStaR Institutional Repository project at Cornell (http://datastar.mannlib.cornell.edu).

While online education has certainly been very different than my past campus-based experiences, I have found the environment to be effective and the classes rigorous to say the least! It was fortunate my university is currently offering a program that allows me to drop my time to 50% time without penalty in terms of my benefits. It is also fortunate that the groups I work with encourage continuing learning. Even with everyone's good wishes, though, it's not a simple task to juggle my schoolwork with my job. I've learned a lot about budgeting time and am still exploring ways to schedule effectively. Some of the other students I've met in class have made other arrangements. Some don't work at all while others reduce class loads to work full time, Local Illinois students have the option of taking a GA (graduate assistant) position. My classmates have a variety of backgrounds, including library science, finance, wireless communications, law, Luso-Afro-Brazilian studies, computer science, and biological science. The age range of my classmates is quite varied, with some just having graduated college and others nearing retirement.

After considering programs at the University of Washington and the University of California, Los Angeles, I chose the UIUC iSchool for a number of reasons including a science-specific data curation track, the synchronous classes and on-campus component required for distance students, the reputation of the school and the faculty, and the number of faculty researching topics of interest to me. However, there are many other options for those with different paths in mind! Many of the schools offering distance programs have continuing education classes; at UIUC, graduate classes offered to enrolled MLIS students are open to anyone with a bachelor's degree (as 'community credit students') for continuing education depending on space availability (http://www.lis.illinois.edu/programs/cpd/). At the other end of the academic spectrum, options include Doctoral degrees through integrative programs such as Science Studies through University of California San Diego (http://sciencestudies.ucsd.edu). Interestingly, while there are not so many telecommunting opportunities at the undergraduate or PhD levels, it seems that a variety of distance learning program options exist at the Masters level. I am happy to share my experience if it would be of interest, please feel free to send along any questions!

News Bits

GIS Working Group Report

edit

Theresa Valentine (AND) and Jamie Hollingsworth (BNZ)

The GIS Working Group met in a breakout session during the Information Managers meeting and covered the following: GIS Standards for LTER Sites, LTERMaps demo, and GIS/Remote Sensing training needs. Other topics were briefly addressed.

The GIS Working Group discussed the process for setting GIS Standards for all LTER Sites. The group felt that there should be a base level of data that each site should be required to make available on their website. The working group plans on developing a formal proposal and presenting it to IM Exec and NISAC. There was a lot of discussion during this session and under the governance session on how to proceed with the requirements. We also discussed the need for having the metadata of this basic level data searchable with other site databases, and documenting the provenance of the data. A centralized portal for search, display, and access for the site GIS data should be developed.

Recommend list of basic level spatial data with associated metadata that each site will be required to have online are as follows:

1. Location data for the site headquarters (input into SiteDB)
2. Boundary polygon or research area(s), and a boundary of Interest for Site (extending outside the research area for DEMs, imagery, etc).
3. Digital Elevation Model (DEM)
4. Imagery (Landsat TM or better scene with multiple years for comparison)
5. Research plot locations of core research plots (based on GPS)
6. Transportation network
7. Hydrography
8. Structures
9. Demographics: Historical demographic data for LTER sites have been developed through the ecoTrends project and are available at the Coweeta site or through a GIS web service developed at the Andrews LTER.

The LTERMaps Team provided a live demonstration of the project. LTERMaps is a cross-site effort between Andrews, Konza Prairie, Bonanza, Georgia Coastal, and Baltimore LTER sites. The project team began meeting through video teleconference calls in January of 2009 to review options for internet mapping and to develop a prototype for Phase 1 of the project. The Phase 1 prototype is an application that can be imbedded into a webpage using the Google MapsTM API. The advantages to this approach are the reliability and performance of Google's satellite imagery, its familiarity to the public, lack of cost, and large user community. In addition, it will be relatively simple to add additional sites (for example, ILTER sites and NEON sites), and individual LTER sites will not be responsible for managing basic satellite imagery. Other spatial data (for example, site boundaries and climate stations) can be added to the Google base imagery for a customized view of the sites, and users can download the GIS formatted files from the webpage. Here is link to the demo site: http://gcelter.marsci.uga.edu/public/gis/LTERmaps.html

The group discussed GIS/Remote Sensing Training needs for LTER Sites and identified a need to conduct a survey of web mapping tools and skills at each site. The following training needs were identified:

1. GIS Data discovery training: how to find data and evaluate for appropriate use.
2. Higher end GIS training: (server and internet mapping applications, training that would work toward GIS Certification for site GIS personnel, modeling, and using the new Flex and Silverlight APIs)
3. Matching GPS data with remote sensing data (including LiDAR).
4. Google Maps training with IM's and using LTERmaps to help sites imbed the technology in their web pages.The group would like to propose a workshop during winter of 2010 (post ASM proposal has been submitted).

Other topics:

1. Are there topics that the group should identify as GIS best practices? Suggestions were data transfer requirements and metadata.
2. Spatial data workbench? What is the status of the project? A website was found with information: (http://www.lternet.edu/technology/sdw/)
3. Barbara Nolen (JRN) agreed to track the Landuse/Landcover efforts that are evolving within the LTER Network.
4. Theresa Valentine (AND) to coordinate with emerging Spatial Analysis Committee (http://asm.lternet.edu/2009/workgroups/lter-remote-sensing-data-information-and-coordination)

GIS Working Group members participated in several other ASM workshops on topics such as LiDAR, Integrating spatial and temporal data across the LTER nework, visualization, Maps and Locals, and land fragmentation.

IMC Governance Working Group

edit

Information Management Committee (IMC) Meeting September 2009

The discussion of governance focused on decisions made by and communications between committees and boards across the LTER organization framework. A figure developed by the IMC GWG over the last year was shown an aid to facilitate discussion. Some current roles and practices were discussed:

- LNO recommends when working groups or sites have needs for resources and implementation that they contact NISAC as a way to approach the EB.
- Some NIS modules are developed by sites, while some are a whole network effort (PASTA); NISAC could play a larger role in reviewing and coordinating this.
- We need to be aware as we move from informal to semi-formal to formal in our practices, the strengths inherent to the informal/semiformal and that exploring similarities and differences make for stronger products
- A critical mass of sites may be useful for vetting
- Science working group projects go through an approval process with the EB, while LNO holds funds for the IMC, but how are projects funded? For now projects may be funded outside the IM. It would be useful for IMexec to have an annual budget so they can plan accordingly.
- Form of affiliation for projects may be useful to have, each with its' own criteria, a designation as to their relationship with LTER
  - LTER affiliated
  - LTER Endorsed

  and if part of LTER they have a designation as to their development when related to NIS, ie

  - NIS core, in production
  - NIS core, in development
  - NIS core, proposed
  - NIS module, in production
  - NIS module, in development
  - NIS module, proposed

The following link is to a diagram that was used as the basis for an initial discussion.

http://databits.lternet.edu/sites/databits.lternet.edu/files/IMC_gov.pdf

Please note that this diagram may be missing elements as the types of decisions and who makes them was not discussed in detail at the meeting or at any time by larger groups of information managers. However a need was identified to improve procedures for knowing who needs to be making decisions. Alternate models may need to be explored.

We focused on how IMC and IMexec interfaces with NISAC and the EB, as well as the LNO to facilitate and support progress on potential NIS modules, standards, and other projects related to IM. NISAC is well positioned to broker communication between IMexec and EB. The work of the controlled vocabulary working group was used as an example to explore not just who, but how decisions are made regarding the design, development, testing, support and adoption of IM or science driven CI "products". Working groups that have an idea or design for a product should speak with NISAC early on to get people thinking about how products can be developed to fit within a coordinated CI LTER NIS framework. This correspondence can happen in the form of a short report including the current status of the effort, useful ways it may be implemented, and some technical specifications. NISAC can use this information to propose how such tools may be incorporated into other projects. It was recognized that we have moved from developing tools, standards, and best practices in informal ways, to using semi-formal consensus to advance efforts, and now may need more formal steps. During the design and development phase of any project NISAC may work closely with IMC, working groups, IMexec and LNO. At the point when resources are assigned from LNO a more formal process should be followed between NISAC and the EB.

We also discussed the partnership between IMC and LNO and how LNO resources are commonly available when LNO staff are serving as active participants and developers on working groups. In 2004, NISAC developed some requirements for potential NIS modules. It was recommended that this document should be revised. In addition, since 2004 network governance has changed so that EB now would now assign resources through LNO. This arrangement, however, means that technical proposals are not vetted through a group with the appropriate expertise. This procedure needs to be revised to handle proposals in a more informed way so the scope and ramifications of the project as well as its coordination with the larger network framework is taken into account. This may mean that in addition to the traditional flow of information from NISAC to EB that EB will ask for a review or advice from NISAC on issues relating to NIS and data. In addition there was wide agreement that potential NIS modules and other projects related to IM standards, protocols, best practices, and controls should address in written form a set of topics including but not limited to design, prototyping, testing, maintenance, update, site-level enactment, training, etc.

LTER Unit Registry: Products and Processes

edit

Mason Kortz (PAL/CCE)

The LTER Unit Registry is an online database designed to help information managers and other LTER users query, create, and maintain lists of scientific units. With a scheduled launch of spring 2010, it seems like a good time to start talking about how users will interact with the registry. Currently, we have two products planned for launch. The first is the Unit Registry Web Service, a development tool targeting users designing their own unit-aware applications. The second is the Unit Registry Interface, a web site that allows users to interact with the registry through a simple graphic user interface.

Unit Registry Web Service

The Unit Registry Web Service is a REST service currently in development by the Unit Working Group. This service accepts HTTP requests for information about scientific units and quantities and their usage by LTER sites. Requests may be made for all units, a single unit using a unique identifying number, or a set of units using search filters. The service returns information in an XML or JSON document, depending on the type of request sent.

The Unit Registry Web Service allows the registry to be used in community developed applications at both the site and network level. The service provides a tool similar to an API, but with the advantage of centralized data storage and maintenance. This provides developers with access to the contents of the Unit Dictionary and the query functionality of the Registry. The service can be used in any development environment that allows opening of files over HTTP, including C, PHP, Perl, Java and most other programming languages, as well as analysis tools such as Matlab and R.

The Unit Registry Web Service is intended for use by information managers who are developing data systems and tools that require scientific units. In comparison to a site-specific unit database, the service reduces development time by providing functionality and promotes standardization by serving a single body of content. As a centralized development tool rather than a centralized application, the service allows sites to develop solutions for local needs while leveraging the combined content of the LTER network.

Unit Registry Interface

The Unit Registry Interface is a web site that presents a graphic user interface to the Unit Registry. It provides the same search capabilities as the web service, but inputs are controlled by web forms instead of by HTTP requests, and the results are displayed in human-readable format. The form elements on the site permit multiple query results to interact with each other, allowing advanced options such as searching within previous results or viewing the union of two result sets. Search results can then be downloaded in common formats such as CSV, Excel, or XML.

The Unit Registry Interface is one use case of the Unit Registry Web Service. Thus the inputs and outputs of both interfaces[??], while different in presentation, access the same body of content. Developers using the web service can use the interface to verify query results. Also, the code for the interface is available via subversion from the LTER Network Office servers, so developers can use the Unit Registry Interface as a working example for how to create a web application that uses the Unit Registry Web Service. The Unit Registry Interface is designed for information managers, researchers, and members of the public who want to use the registry as an application rather than a development tool. As a centralized application, it allows access to the units used by all LTER sites in a user-friendly way. This is intended to meet the most common query and download needs of LTER users without requiring extra development time at the site level.

More information about the Unit Registry can be found on the LTER IM forums at: http://intranet.lternet.edu/im/forum/7

Developing a Controlled Vocabulary for LTER Data

edit

John Porter, VCR/LTER

During the one-day 2009 LTER Information Managers Meeting, there was a working session on the creation of a controlled vocabulary for use in providing a more consistent keywording for LTER documents. Past analyses had indicated that LTER were highly esoteric, with over half the keywords used only a single time (http://intranet.lternet.edu/archives/documents/Newsletters/DataBits/06sp...).

The working group presented a draft science keyword list derived using the following procedure:
Inigo San Gil and Duane Costa assembled list of LTER EML keywords from the LTER Metacat, that list was then cross-linked to the NBII Thesaurus (http://thesaurus.nbii.gov), the Global Change Master Directory (GCMD) keywords (http://globalchange.nasa.gov/Resources/valids/) and words recently used by Metacat searches. Based on ANSI/NISO Z39.19-2005 (Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies) standard, words on the list were changed to preferred forms and a list of synonyms created. Specific place names and taxonomic names were removed, because they can be handled more succinctly in other parts of EML documents, or in alternative keyword lists.

Keywords were then selected from the list based on two criteria. Any word that was used in the NBII and GCMD word lists was automatically included, and any word that was used by two or more different LTER sites was included. This draft list was then circulated to LTER Information Managers who suggested additions and deletions. A SurveyMonkey survey was used to get "votes" for or against specific words and changes were made to the list based on the results of the survey.

The final list has 640 keywords, with an additional 148 synonyms (including the non-preferred forms of words used in the list). Thirty-one percent (201) of the keywords were also found in the NBII Thesaurus and 21 words were found in the GCMD keyword list.

Following the report, discussions focused on several issues. The first was what steps needed to be taken to make the final list an "official" LTER list. The sense was that such a step needed to be taken by the LTER Executive Board (EB) and the working group was charged with preparing a proposal to the EB for adoption and on procedures for managing the list.

There was also substantial discussion of other efforts (e.g., SONET, Semtools) that might be leveraged. The group discussed the need to develop tools such as autocomplete keywording tools (similar to the one implemented by Duane Costa on the LTER Metacat to suggest search words) that could be deployed at sites to help in document creation, and semi-automated tools that could analyze document content and suggest words (i.e., semantic annotation).

Also discussed was leveraging the keyword list by creating hierarchical polytaxonomys which link keywords to more general concepts, and Barbara Benson reported on preliminary steps along these lines. Additionally, the group suggested that augmentation of the synonym list, so that a wider array of word-forms could be linked to the preferred list, was desirable. There was also discussion of some innovative ways the word list could be used, such as identifying "unfindable" datasets that had no keywords drawn from the preferred list.

Presentations and notes from the working group, along with the list can be found at: http://intranet.lternet.edu/im/node/489.

Beware the Ida's of November

edit

John Porter (VCR)

In mid-November 2009, extratropical storm (formerly hurricane) Ida visited the VCR/LTER bringing with it sustained high winds and tides. During the storm we were able to monitor a variety of conditions on isolated Hog Island (20 km from our lab) using our wireless network. See graphs and webcam images in the attached PDF file.

How serious was Ida? In Norfolk, just to the south of our site, it set a new record for storm surge – exceeding both Hurricane Isabel (2003) and two major hurricanes in 1933, and came in fourth for overall flooding (surge plus tide) (http://www.wunderground.com/blog/JeffMasters/comment.html?entrynum=1383). On Hog Island, winds on the top of the Machipongo Station (an old Coast Guard Station, now run by The Nature Conservancy), sustained winds of more than 50 mph were observed. On the ground, and surrounded by vegetation, winds were more moderate, topping out at about 25 mph (11 m/s). However, an important aspect of Ida was its duration. Most nor'easters have a duration of 1-2 days, but Ida had sustained winds for more than 4 days in a row – all coming from approximately the north. This led Ida to pile up water in the coastal bays resulting in extremely high tides over an extended period. This in turn led to extremely high ground-water levels as the less-dense freshwater floated on top of the tidally-driven salt water.

As several LTER sites have found, extreme events often provide novel stresses on monitoring equipment. Unlike Hurricane Isabel in 2003, when we lost several stations to flooding, almost all our equipment stayed dry enough to function throughout the storm. Our closest call came at a flux tower located on a salt marsh. We had recently raised all the equipment an additional 45 cm due to concerns about storm flooding. Nonetheless, the outer equipment box did flood, but the internal NEMA boxes safely floated, preserving expensive instruments, computers and data loggers. A solar controller was less fortunate – its NEMA container was wedged so it couldn't float and salt water caused a small, electrical fire that destroyed the controller. We also had some problems with a tide sensor, when an intermediate junction box partially flooded, causing a several day outage. One webcam also had problems during the height of the storm, but returned to normal operation following the storm, allowing a rapid assessment of overwash in the vicinity of the camera. However, our other main webcam stayed up throughout the storm.

How do you put your GIS metadata in EML format?

edit

Jonathan Walsh (BES)

A quick email poll was conducted to see how we turn GIS metadata into EML.

13 responses were received, which represents half the LTER sites.

The results were as follows:

- Have GIS metadata but not in EML: 4
- Do not have GIS data: 3
- Use XML transform: 3
- Use a script: 1
- Paste in manually: 1
- Morpho: 1

Good Tools And Programs

FireBug: Web Customizing To Fit Your Needs

edit

Sean Wiley (PAL/CCE)

One of the must-have Firefox extensions for web developers is Firebug.

Firebug is most commonly used as a debugging tool because of its ability to facilitate alteration of web pages in 'real time'.

However, another aspect of this tool is its ability to modify web pages for personal use. In taking on the task of converting a webpage into, in this specific case, a desired pdf form, one issue that arose was the inability to modify certain fields integrated into the template, i.e. text fields, search/tool bars, jpeg images, etc.

With the versatility of Firebug, you harness the power of modifying all aspects of a webpage into a desired form before saving it in a particular format.

The most valuable aspect of Firebug is that as you change the HTML/CSS/JavaScript code, the changes are reflected immediately in the browser. Because of this "safe exploration", FireBug is a tool that can be used by novices without much knowledge of the code as well as advanced programmers looking to implement an algorithm on the webpage.

FireBug offers a diversity of development tools to the user; it can be used as not only a debugger, but also as a general-purpose modifier.

Matplotlib: An Open Source Python 2-D Plotting Library

edit

James Conners, CCE / PAL

There are plenty of libraries available for producing plots of data. Some are built into proprietary application environments like Matlab (http://www.mathworks.com/) or offered as a service on the web like Google Charts (http://code.google.com/apis/chart/). The majority of others are written in a particular programming language over pre-existing or custom graphic and other supporting libraries, like PLplot (http://plplot.sourceforge.net/), Gnuplot (http://www.gnuplot.info/) or JPGraph (http://www.aditus.nu/jpgraph/). Matplotlib is in the last group of libraries, written in Python and dependent primarily on the Numpy (http://numpy.scipy.org/) scientific computing package. It's open source and maintained by an active community of developers and users.

After using JPGraph for about three years to produce dynamic plots of data on our web sites, we've recently switched to Matplotlib. We made the change to an implementation using Matplotlib because of the greater variety of plots available and the improved performance with large datasets. One of our first activities was to build a web service using a set of plotting classes abstracting the library's application interface. Hooking the library's functionality into a web architecture is a pretty simple process using either the mod_python (http://www.modpython.org/) Apache module or Python's CGI interface. Another way to use Matplotlib is within an interactive shell – such as the built-in Python interpretor or a more advanced one like iPython (http://ipython.scoi_events.xlsxipy.org/moin/). These shells allow you to create and manipulate plots using the pylab interface, a procedural set of functions built over the Matplotlib API that feels like the Matlab graphics interface. The higher-level programming interface works well for quicker implementations while the extensive object-oriented API provides the flexibility for building more customizable plotting routines. There are also additional toolkits available for download that supplement the library with capabilities like 3-D graphics and plotting over maps.

There are always going to be cases where your preferred plotting library or application is either inadequate or comes with too much overhead for what you need. From our experience so far with Matplotlib, the library seems to strike a good balance in being flexible enough to handle the majority of visualization tasks while still being a good choice for small one-time tasks that require a quick view into the data. Since the community of developers and users seems to be stable, if not growing, there has recently been more documentation available for help on getting started with the library, as well as a new book (see Resources) recently published, easing the learning curve a bit. At this time, it fits our needs as a quality plotting library with an active development and support trajectory.

Resources:

Matplotlib home page: The site contains documentation, galleries, user-contributed how-to's, etc.
[http://matplotlib.sourceforge.net/]

Matplotlib book:
[http://www.amazon.com/Matplotlib-Python-Developers-Sandro-Tosi/sim/1847197906/2]

SciPy Matplotlib Cookbook: Quick examples to follow when getting started making plots
[http://www.scipy.org/Cookbook/Matplotlib]

Matplotlib mailing lists:
[http://sourceforge.net/mail/?group_id=80706]

Video Lecture given by the Matplotlib creator:
[http://videolectures.net/mloss08_hunter_mat/]

Good Reads

A book for data people

edit

Jonathan Walsh (BES)

The Fourth Paradigm: Data-Intensive Scientific Discovery is a collection of essays about Jim Gray.

Jim Gray gained fame as a database innovator. His contributions to database software development include transaction processing.

He disappeared at sea three years ago.

Gray asserted that computing is transforming how science works. Scientists will be faced with more data than they can handle and in order to prevent overload a new suite of computing and visualization tools will have to be developed. He named this development a "fourth paradigm".

For more on this see New York Times Science Edition December 14 2009 online here:

http://www.nytimes.com/2009/12/15/science/15books.html?_r=1&hpw

Nature special issue on Data Sharing

edit

John Porter VCR/LTER

http://www.nature.com/news/specials/datasharing/index.html

In September 2009 the prestigious journal "Nature" published a special issue on data sharing that shows that ecology is not the only discipline to be addressing this topic.

For the opening editorial, the title tells it all: "Data's shameful neglect." The editorial makes the forceful point that "Research cannot flourish if data are not preserved and made accessible. All concerned must act accordingly." Ecology is not alone in having growing pains. The editor notes across science: "All but a handful of disciplines still lack the technical, institutional and cultural frameworks required to support such open data access." The editorial calls for funding support, and new roles for agencies and the scientific community in managing data. This will require new efforts to educate scientists - as the article concludes: "data management should be woven into every course in science, as one of the foundations of knowledge."

This is followed by a feature article on "Data sharing: Empty archives," which addresses the challenges that have confronted efforts to create archives of shared data. The article includes discussion of the new "DataOne" project being headed by Bill Michener.

The feature is then followed by two opinion pieces derived from international workshops on "Prepublication data sharing" which provides recommendations and a rationale for sharing data even prior to publication, and "Post-publication sharing of data and tools" which advocates creating a "research commons" to help overcome proprietary and licensing issues.

Thanks to Margaret O'Brien and James Brunt for bringing the issue to my attention.

Standards and Data Sharing in Ecology

edit

John Porter, VCR/LTER

Review: New Knowledge from Old Data: The Role of Standards in the Sharing and Reuse of Ecological Data by Ann S. Zimmerman, Science Technology Human Values 33:631. 2008, DOI:10.1177/0162243907306704

This article provides an "outsider view" of data sharing in ecology. Author Ann S. Zimmerman is a librarian specializing in data and data sharing, and has a long history with LTER and NEON. The article addresses topics such as "Data and Data Reuse," "Standards as Distance Spanners," "Ecological Data and the Practice of Ecology," "Locating and Acquiring Data," "The Relationship of Ecological Identity to Reuse", "Recognizing the Importance of Purpose," "Dealing with Uncertainty," "Focusing on the Object of Study," and "Focusing on the Data Collector."

With such a list of topics, it's no big jump to conclude that this article will be a "good read" for LTER Information Managers, especially in a time where there is increased emphasis on data integration and synthesis. Dr. Zimmerman provides the context for the study of standards and draws on interviews with ecologists to help us understand the roles they play in the reuse of ecological data.

The article documents how a variety of ecologists (e.g., modelers, foresters, plant, aquatic and avian ecologists) reused data for articles in Ecological Society of America journals, and was based on interviews during 2001-2002. Most of the data being reused came from tables and graphs in publications, or direct contact with researchers, although a few were based on publicly-available datasets. Interestingly, she found that "The ability to understand data was the most important factor for reuse, and it was this requirement that binds together ecologists' experiences." That is, ecologists tended to use types of data that they were most familiar with. This links to her fundamental thesis that experience and training provides a context for understanding data that extends beyond the formal metadata available. In specific she found that: "Their experiences in the field or laboratory, in combination with formal disciplinary knowledge, provided ecologists with the expertise to understand the critical link between research purpose, methods, and data; to recognize the limitations of particular types of data; and to visualize potential points of data collection error."

With regard to standards, she found that there were a wide variety of factors that LTER considered when choosing a methodology, such as the purpose of the study, and the cost and availability of instrumentation. Her conclusion was that "There are legitimate reasons for the use of different research methods, which helps explain why the ecologists I interviewed did not place overriding emphases on methodological standardization."

I found the article to be most illuminating for the perspective of "point-by-point" syntheses - where a few data values are drawn from a wide variety of sources (e.g., different publications). It would be interesting to see additional treatments of this subject now that there is a greater degree of data availability and metadata standardization. Nonetheless, the article provides clear documentation of many of the factors that influence ecologists when reusing data.

Identifying Best Practice and Skill for Workforce Development in Data Curation

edit

Karen S. Baker (PAL/CCE)

P. Heidor and H.Tibbo, 2007. Identifying Best Practices and Skills for Workforce Development in Data Curation. American Society of Information Science and Technology ASIST (http://www.asis.org/Conferences/AM07/panels/41.html)

What's in a name?

This good read allows us to consider the role of the information manager together with those of data curator and data repository director. A panel at the American Society of Information Science and Technology (ASIS&T) annual conference in 2007 explored a topic that represents an elaboration on one facet of the LTER information manager's work with datasets supporting research activities, namely, the role of data curator of a data repository. As the emphasis on data access, integration, and interoperability continues to grow, so do expectations for those in the role of information manager. It seems useful to review the panel abstract as it represents an emergent awareness that informs standards-making, requirements-setting, and assessment-enabling efforts. The recent proposal preparation document reminds us that "One of the recognized strengths and pioneering aspects of the LTER network relates to information management and technology". For LTER, the pioneers, in creating both a place and expectations for a new role, named this role 'data management'. We might reflect upon how in the 21st century LTER broadened the data management role to that of information management while other communities are beginning to articulate the need for data curators. One might say, it's all to do with the web-of-repositories (Baker and Yarmey, 2009; Baker and Millerand, in press).

Panel Abstract:

The nature of science and scholarship is being transformed by the ability to collect and integrate vast quantities of information. Some sciences such as ecology and environmental science are inherently integrative, requiring the combination of information of many types from many sources in order to answer more complex questions than previously possible. This new information and the information management tools designed to deal with this volume of data will help us make informed decisions that will impact human health and prosperity. To enable this cross-scale, interdisciplinary integration for the coming generations of scholars, data must be managed to facilitate interoperability, preservation, and sharing. We define this discipline of "data curation" as the practice of collection, annotation, conditioning and preservation of data for both current and future use. Government and industry have recognized both the opportunities and challenges and have called for improved data curation. Current data curation challenges can be grouped into two classes: underdeveloped data curation practice and shortages of skilled data curators. In this panel we will explore methods to maximize our opportunities and the impact of data on scholarly work. This discussion will be lead by panelists informed by government studies, successful practice and curriculum development projects. The primary questions to be addressed are: what are the required skills for data curation, who should learn these skills, how do these individuals fit into the social fabric of science and who pays for this new work?

The Ghost Map by Steven Johnson

edit

Theresa Valentine (AND)

"This is a story with four protagonists: a deadly bacterium, a vast city, and two gifted but very different men."

So begins Steven Johnson's multi-layered account of the 1854 London cholera epidemic.

London was just emerging as one of the first modern cities in the world, but it lacked the public health infrastructure to support its exploding population.

As a result, the city became the perfect breeding ground for a deadly disease.

Rising up against the dogma of the scientific community, two men, Dr. John Snow and Reverend Henry Whitehead, attempted to put a stop to the epidemic, and in doing so revolutionized the way we think of the spread of disease, the nature of scientific inquiry, and the rise of the modern city.

Calendar

Events

edit

Event: International Conference on Scientific and Statistical Database Management

Location: Heidelberg Germany
Dates: June 30-July 2, 2010
Web: http://www.ssdbm2010.org/

SSDBM provides a forum for original research contributions and practical system design, implementation and evaluation. Individual themes differ year to year with the main focus remaining on databases and application in the scientific and statistical fields. Recent themes included geospatial and sensor databases, bioinformatics (genomics, biodiversity informatics including biological databases), geological databases, data mining and analysis, metadata management, conceptual models, data integration, information visualization, scientific workflows, and system architectures.

Event: MULTICONF-10

Location: Orlando, Florida, USA
Dates: July 12-14, 2010
Web: http://www.promoteresearch.org/

The primary goal of MULTICONF is to promote research and developmental activities in computer science, information technology, control engineering, and related fields. Another goal is to promote the dissemination of research to a multidisciplinary audience and to facilitate communication among researchers, developers, practitioners in different fields.

Event: BIOCOMP'10

Location: Las Vegas, Nevada, USA
Dates: July 12-15, 2010
Web: http://www.world-academy-of-science.org/worldcomp10/ws/conferences/bioco...
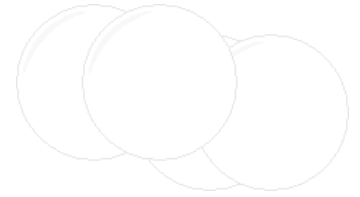
Event: Ecological Society of America Meeting

Location: Pittsburgh, PA, USA
Dates: August 1-6, 2010
Web: http://www.esa.org/pittsburgh

The Ecological Society of America will place global warming at center stage to draw a critical combination of scientists, policy makers, and concerned citizens to understand further its causes and consequences and to elucidate a clear scenario for addressing what is perhaps the most serious environmental threat facing the biosphere.

Theme by Dr. Radut.