LTER

LTER Databits

Information Management Newsletter of the Long Term Ecological Research Network

01001100 01010100 01000101 01010010

Spring 2014

Welcome to the Spring 2014 Issue of Databits!

This issue is dominated by two major themes - the past and the future. As we experience progressive changes in information management practices and technologies there is an illusory sameness, a sense that what "is now" always was and always will be. However, as articles in this issue will show, nothing could be further than the truth. There is virtually no aspect of the way LTER manages information that has not undergone dramatic changes - and that will not change again in the future (which is what makes Information Management challenging and fun).

For the past, we have an extensive review of the development of geographical resources in the LTER Network, a retrospective on the career of Susan Stafford, and a rough timeline extending from the start of LTER to the future.

This issue is also not without the present. A new resource for information on managing sensors, approaches to integrating data, and a survey of how some sites manage bibliographic data are all discussed. Additionally there are reviews of recent articles of interest in the Good Reads section.

Most dangerous, given the oft-quoted (but difficult to attribute) statement "predictions are difficult- especially about the future," we have some commentaries that try to dust off a crystal ball, with discussion of how data will be archived in the future, and James Brunt's predictions about how LTER Information Management will radically change in the future (as it has in the past).

Co-Editors: John Porter (VCR) and Mary Martin (HBR)

Featured Articles

Commentary

Good Tools And Programs

Good Reads

Featured Articles

History of Geographic Information Systems (GIS) in the LTER Network

edit

Theresa Valentine (AND)

GIS use started growing within the LTER Network in 1989 after a report to NSF (Shugart Committee Report) and the development of the Minimum Standard Installation (MSI) resulted in supplemental funding to LTER sites, where the majority of the proposals sought to develop site competency in the areas of GIS and connectivity. There was a series of workshops that were funded by NSF, one a GIS training and the other a Remote Sensing Workshop. Sites acquired GIS software and hardware (mostly UNIX boxes with color screens, digitizer, plotter, and ARC/INFO software). The GIS and computer industry were in transition from mini-computers to UNIX systems.

In 1990 a survey was conducted to provide a snapshot of the technologies used for information management at each site. This survey was used to develop a report (Foster and Boose) to NSF on the effectiveness of the supplemental funding program.

GIS was a hot topic with information managers, based on a review of the early DataBits articles and IMC meeting notes. There was a GIS Corner column in the newsletter, and additional articles on issues around GIS and UNIX administration. During the early 1990’s, GIS and computer technology continued to evolve quickly, with increased computing power on personal computers, and the development of PC ARC/INFO and user interface programs such as ArcView. Plotting capability went from pen plotters, where shading was difficult, to the development of HP InkJet plotters. Costs for hardware and software were decreasing and universities started offering classes in GIS and automated cartography. Access to GIS technology moved from working in a laboratory setting, to desktop

accessibility.

Many of the LTER sites cooperated with Universities in setting up GIS laboratories, with digitizers, plotters, computer terminals, and software for GIS and Remote Sensing capabilities. The Network Office, located in Seattle, had a GIS/Remote Sensing computer lab that site researchers could access remotely or locally.
GPS technology was becoming available, although the cost of GPS units was high, and satellite access was somewhat restricted. The Network purchased several GPS units, and distributed them around the sites, and the technology and expertise was shared. Sites would combine training with field campaigns to gather GPS locations for study sites and site infrastructure.

The All Scientists Meeting (AMS) has been a forum for discussion, sharing on GIS technology, spatial data management, and research or science topics related to spatial analysis. The 1990 ASM had workshops on GIS, remote sensing, and GPS.

The 1994 Information Managers meeting in Seattle stands out as a critical point, where open sessions were held on the management of spatial data and inter-site data access. There were invited speakers with breakout sessions. Groups worked on spatial metadata standards, spatial data exchange, user access and image catalogs, and proprietary issues. Recommendations included defining the capabilities of a spatial, visual, statistical, and modeling lab. “This should include GIS/RS capabilities, a wide range of statistical and visualization software and a combination of hardware, software, and technical personnel necessary to promote ecological modeling.” Nancy Tosta presented on the Federal Geographic Data Committee (FGDC) and the National Spatial Data Infrastructure. This was the beginning of the FGCD geospatial metadata standard, and metadata as something LTER should be thinking about.

Something interesting happen after this meeting, as GIS fell off the radar, at least at the network level. Focus for the Information Managers switched to developing the Network Information System (NIS), the Network Office moved to University of New Mexico, Jerry Franklin retired as chair of LTER, and the DataBits newsletter went on a hiatus when John Porter left for a detail to NSF. The next time GIS is mentioned is in the 1999 report of the LTER Technology Committee. The Technology Committee was working to acquire Landsat data for LTER sites, and was working with the San Diego Super Computing Center, focused on high performance computing, data storage, modeling, and visualization of ecological data.

LTER Information Management Committee GIS Working Group Formed

LTER GIS picked up again at the 2000 ASM in Snow Bird, Utah. Ned Gardner organized a workshop called “GIS on the Internet and LTER: A frontier for research, applications, and solutions.” The IM GIS Working Group was formed from this workshop to develop cross-site and intra-site research proposals that would help push internet mapping technologies in directions appropriate to ecological research. Early members of this group included Ned Gardner, Theresa Valentine, Barbara Nolen, Todd Ackerman, Peter McCartney, John Vande Castle, and Ken Ramsey.

Several workshops on Internet mapping and web services were held jointly with the San Diego Super Computing Center, with representatives from LTER sites, universities, and the private sector. The big topics were Internet mapping, spatial data and EML, the LTER Spatial Data Workbench, web services, spatial and non-spatial data integration, and delivery of data to researchers. The group conducted a survey to find out what types of web mapping sites were doing, and in 2003, a GIS/Internet Mapping edition of DataBits was produced (Fall 2003). It is interesting to note that an “editorial” section in DataBits was developed after this somewhat controversial issue.
GIS software was now established on the PC desktop, and emerging technologies included managing spatial data within a relational database structure through ArcSDE or Oracle Spatial, ArcGIS with extensive desktop GIS tools in a GUI (command line and macro tools were gone), Internet mapping programs (ArcIMS and Minnesota Map Server). The 2002 Winter IMExec meeting notes in the Spring 2002 DataBits edition recommend that Spatial data, traditionally managed by GIS and remote sensing experts, should be incorporated into IM systems at the site and network level.

The GIS Working Group focused their interactions on workshops at the All Scientists Meetings, with little interaction between, although there was some when GIS folks attended the annual IMC meetings. Workshops at the ASM in 2003 and 2006 were successful, but it was difficult to sustain involvement in between meetings.

GIS technology has moved to the desktop and the internet, with mapping applications such as Google Maps and Google Earth, the availability of world wide data sets through map services, and the increased proliferation of inexpensive GPS units. GPS satellites became more abundant and available to non-military users. Phones were geo-referenced and citizens became more spatially aware with MapQuest, and navigation programs for vehicles. This is the start of the downsizing of computer equipment, with smart phones and tablets.

A GIS survey was conducted in 2007 and identified a contact for each site, and what base layers were available, information on status of digital ortho-photography, Landsat scenes, core site locations, EML, spatial data on-line, and internet mapping capability. At that time only 8 sites did not have Internet mapping capability, most sites had digital elevation models, and all sites were using ESRI products. This survey was conducted by Barbara Nolen (JRN). The 2007 IM Committee meeting in San Jose, California had a town hall discussion on integrating diverse data types and specifically spatial data into the LTER cyber infrastructure. Mike Goodchild from UC Santa Barbara addressed the challenges of managing and utilizing spatially enabled data. The GIS working group updated recommendations for GIS at sites. Recommendations included basic spatial information and directions for geo-spatial referencing for study locations. The idea about developing a centralized portal for users to search, display, and access all site spatial data and a strategy for providing basic map service capacity for every site in the network was formulated. The group met for two planning sessions in 2008, and made formal recommendations to the IMC. The idea of a centralized portal for access to LTER GIS data was brought up again. A core GIS team worked together over the next year, meeting via teleconferencing, and produced a prototype to be hosted on the Georgia Coastal Website. Thus, LTERMapS was born.

The first LTERMapS internet mapping workshop was held in March of 2010, as a post-ASM product-oriented workshop. The workshop allowed several of the LTER Information Managers to work on porting the project to the Network Office and connect directly to the SiteDB database. The group completed a guide for using the Google Maps API, and scoped out specifications for the next phase of the project. Key players included Jamie Hollingsworth, Adam Skibbe, Theresa Valentine, Jonathan Walsh, and John Carpenter.

The working group continued to work on developing standards for GIS data, and tools for creation of EML from FGCD metadata. Jonathan Walsh and Theresa Valentine participated in a Best Practices for EML workshop, to help with examples of how to document site locations, geographic coverage, and GIS data formats within the EML structure. A workshop on documenting GIS data in EML was held in 2010 at the Andrews Forest. Eleven sites were represented, and new versions of esri2eml style sheets were tested and developed. A document on the best practices for documenting GIS data to LTER standards was developed, along with the specifications and requirements for a centralized service for LTER GIS data called LTERMapS Phase 2. The LTER Cartographic Atlas was developed by Jamie Hollingsworth through an IM time buyout. The Atlas provides cartographic overlays for use in making presentation graphics, across the entire LTER network. Another survey was conducted, to try and get an idea of what types of GIS data sites had and what it would take populate a centralized server.

LTERMapS Phase 2 morphed into GeoNIS, which became a reality through supplemental funding by the LNO for face to face meetings of the team, and for programming support. The group edited the Spring 2012 edition of DataBits, highlighting geospatial activities at both individual and cross-site scales. This edition provided an opportunity to update the LTER community on the cross-site activities.

The GeoNIS project moved forward in 2013 and was developed to interface with PASTA, and provide a way to quality check spatial data in PASTA, and to provide web mapping services for LTER sites. The GeoNIS is currently in beta stage, with automated processes to ingest EML files from PASTA , download data, and create web mapping services for each site. There is an end user application to check for errors, and view the data (http://geonis.lternet.edu/). Key personnel

included Adam Skibbe, Aaron Stephenson, Jamie Hollingsworth, Theresa Valentine, Inigo San Gil, Mark Servilla, Ron Beloin, and Jack Pederson.
Key products and links for the LTER GIS Working Group:

1. LTERMapS (http://www.lternet.edu/sites/map)
2. Documenting Spatial Data to LTER Network Standards (http://im.lternet.edu/project/GIS_document),
3. GeoNIS (http://im.lternet.edu/project/GEONIS)
4. Cartographic atlas (http://www.lternet.edu/sites/lter-interactive-cartographic-atlas).
5. GIS Working Group Terms of Reference (http://im.lternet.edu/sites/im.lternet.edu/files/GIS_working_group_TOR_v2.doc)
6. GIS Working Group project page (http://im.lternet.edu/projects/gis_data)

The Future of GIS in LTER over the next 20 years GIS Technology (also known as Theresa Valentine's crystal ball)

GIS technology is moving towards Open Source solutions, sharing code on sites such as Github (https://github.com/). The industry leaders and large organizations are pushing cloud computing, with software and data located in the cloud, and users working remotely on a variety of lightweight devices. Users should be able to access their data from any device, and applications will have to scale in size to work from large screens to small wrist phones.

Data streams continue to increase and the availability of LiDAR, satellite data, and digital orthophotography are challenges for current computer networks. Moving large data sets and models around in the cloud currently can be cost prohibitive, and it is unclear where the industry is moving in regards to high end GIS computing. ESRI is releasing a ArcGIS for Professionals, however it is in beta release, and it's unclear how it will work with scientific needs, especially with large raster datasets. One should expect that band width and wireless accessibility should increase, although there most likely won't be a business model to cover remote areas, where many LTER sites are located.
GPS units are becoming more plentiful, smaller in size, with increased accuracy, and are now found in phones, tablets, cameras, watches, and glasses. Technology currently exists that will track a person/vehicle as they move between boundaries of services areas (called geo-fencing), sending alerts, notifications and even advertising back to the person/vehicle (http://whatis.techtarget.com/definition/geofencing). This emerging technology might be useful in tracking animal movements.

Plotting of hard copy maps is being replaced by on-line web applications, and large and small format touch screens, with higher definition and 3D capability. 3D printing devices will also change how business is conducted and research tools are developed. Output will be videos, and screen shots.

GIS professionals will be expected to be proficient in computer science, building custom applications for personal devices, with less emphasis on cartography and geography. Data will be available over the web, and staff will need to be good at "mashing" data together from different sources and communicating with researchers to meet their analysis and application need. They will need to be able to sift through the massive amount of data to find relevant information. I suspect that there might be a few hard core GIS nerds left in 20 years, who still like the command line and long for the Workstation ARC/INFO days.

Databases will still be important, as applications will need to capture data from a variety of devices (both researcher and citizen scientist) and provide updated data for the applications. Data will also need to flow seamlessly between different programs, especially with statistical packages and modeling software. Multi dimensional datasets will become more and more important as global climate change continues to be an issue, and researchers will want better tools to visualize and analyze data across several dimensions.

GeoNIS

Future direction of the group is to continue with the GeoNIS development moving from beta to production mode, continued coordination with the IM committee, keeping current on GIS technologies through coordination between site GIS specialists, and renewed commitment of and recruitment of members.

The GeoNIS, when operational, will be the starting point for performing quality checks on spatial data within PASTA. Future development will allow the checks to happen prior to ingestion into PASTA, and additional quality checks will be added. The enhancements on the server side include troubleshooting problems with image datasets, scaling up the server hardware, developing tools for analysis, and keeping up with the changes in the technology, as new software versions are released. Training for site staff would be critical to helping sites and cross-site projects build applications using the web services. There are more tools being added to the ArcGIS Javascript API, and more functionality can be added to the existing application, and other front end applications can be built for specific projects. Additional funding will be needed to finish the project, conduct training, and enhance/create applications.

In twenty years, one would expect that the GeoNIS web services would be the geographic framework for LTER site researchers and cross-site projects to provide access for students, researchers, the general public, and educators to further the goals of the LTER Network and science. The GeoNIS should be transparent to the users, and integrated into the data management and science communities.

Data Integration Experiences

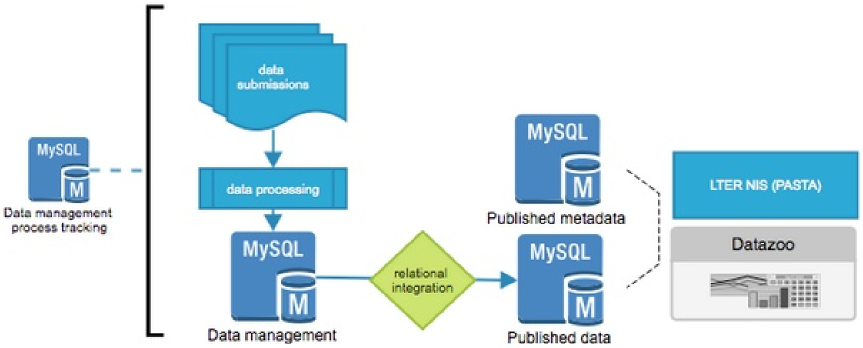
edit

James Connors (CCE, PAL)

Information Management for CCE and PAL LTER has had the opportunity to work on a small set of projects dealing with data integration. These experiences have shaped our current data management workflow and data integration approaches. In general, these experiences have been limited to working to integrate data long after a research plan and data collection have been carried out, so may not be related to data collected according to protocols designed from the outset to support integrated data products.

One of the first approaches to data integration that we implemented was based on simple database table joins. Within our data system interface, a user could select a dataset to begin with (lefthand table) and select another (righthand table) to be joined. The interface provided dropdown selections where the user could create the criteria for joining, specifying which fields in the datasets would be matched on for joining. One of the major criticisms that we received after this implementation was that is was difficult for any users who didn't know the specifics of how the data were related (e.g., which fields served as shared indexes). Also, because the datasets stored in the system are combined across sampling studies, you also needed to be aware of possible differences across studes. Particularities of a specific sampling study that may have affected the indexes and these differences need to be understood to properly assess the quality of the join. In addition to these conceptual problems with the interface's design, which could have probably been addressed, there were additional issues related to storage types. The table joins ran into problems when data storage types for indexes varied across datasets, i.e. strings vs. numeric types, etc. This table joining utility was eventually removed from the system. For our first approach, it gave us insight into a number of specific issues related to integrating data. Foremost, that data integration was more than a technical solution, requiring a level of knowledge of the data that couldn't be generalized.

A subsequent approach to data integration benefited from this experience. As part of our support for the CCE LTER project, a database programmer in our group, Mason Kortz, worked on a project to develop a system for providing integrating data results from two research projects, the Brinton and Townsend Euphausiid database and the Marinovic Euphausiid database. For this project, Mason worked closely with the scientists involved in order to narrow the scope of integration across the two data sources, in addition to defining many parameters that were integrated into the system's functionality in order to accommodate known issues or differences across the data collections. This project produced a functional system (https://oceaninformatics.ucsd.edu/cequi/) that successfully provided integrated data results across projects. The approach taken for this project was the reverse of what was done for the simple table join interface described above. The begin with, the developer worked closely with the scientists to understand the data and to narrow and define the context for integration before a technical approach was designed. The data, it's condition and it's intended use provided the realistic scope that led the design of the system.



Currently, within our primary data system (Datazoo) we are treating the issue of data integration differently. Datazoo serves a wide variety of data across projects and scientific domains. The system provides a set of generalized interface utilities for searching, browsing, downloading, previewing and plotting datasets. Because of this, narrowing the scope and developing a well defined context for integration across the catalogs of datasets is not feasible. With a recent redesign of our data management workflow, a different approach was taken. Relational data across projects is managed in prepublishing databases, where data managers can work with and understand the characteristics of data and how they integrate before they are published to our public system. With this setup we can define integrated products, using database and programming tools, that are then reviewed for quality issues before they are published. The workflow for producing these integrated products is saved and can be rerun after new data are added to datasets. For now, this seems to be the best approach for maintaining data quality as well as a generalized data interface that can accommodate a growing catalog of datasets.

Where are they now? - Susan Stafford reminisces about her years in LTER

edit

Don Henshaw (AND)

Where are they now?

As a means of commemorating the 25th anniversary of the DataBits newsletter, we introduce a new series to remember and reacquaint ourselves with past LTER Information Managers. This new DataBits series is intended to highlight former LTER Information Managers (IMs) by exploring what they recall of LTER meetings, events or other memorable moments during their years of involvement, and update their activities in the following and current years. The intent is to provide an opportunity for any current Information Manager or DataBits reader to contribute profiles of former IMs that they may know and wish to highlight in future DataBits editions.

For this series debut, Susan Stafford graciously agreed to spend an afternoon with Don Henshaw (AND) to reminisce about her years with the Andrews LTER, as Chair of the Information Management Committee (IMC), and succeeding career activities.

Susan Stafford

LTER Site(s): H. J. Andrews Experimental Forest (AND), 1980-1998; Shortgrass Steppe (SGS), 1998-2002

LTER key role(s): Data Management Committee co-chair with Bill Michener, 1982-1993; IMC Chair, 1994-2003

Current status: Faculty, University of Minnesota

(The following is compiled by Don Henshaw but is based on direct quotes and words from Susan with some reorganization and paraphrasing to provide necessary context)

LTER involvement

Andrews LTER: Susan was hired by the Department of Forest Science, College of Forestry at Oregon State University in 1979 as a tenure-track Assistant Professor and consulting statistician, and her arrival at OSU coincided with the original competition of LTER. Susan, as a newly-minted PhD in applied statistics and quantitative ecology from SUNY Environmental Science and Forestry, became the logical choice for the data management leader as the Andrews was selected in the first cohort of six LTER sites. As the Andrews already possessed an enormous resource of past research data from both USFS research and International Biome Program (IBP) funding, Susan quickly recognized the vast opportunity and the need for bridging to these past collections by reorganizing the existing data and establishing protocols and systems to carry forward to the future.

The Andrews LTER IM "team" included both OSU and USFS PNW Station membership and Susan was able to leverage these partnerships and LTER funding to create and build the Quantitative Sciences Group (QSG). The general premise being that what was good for Andrews LTER would be good for the department and College and also good for USFS PNW. The quantitative sciences team was bigger than any one project and was able to facilitate LTER and other research efforts. A science workflow was implemented in the early 1980's where a researcher would consult with Susan on statistical design and defining study objectives, and then proceed to work with other team data managers on collection forms and documentation (the 1980's were the ice-age of data management as the terms "metadata" and "information manager" had not yet been invented). Susan successfully recruited new talent to the QSG as the LTER project's needs grew and expanded into new areas.

Network-level LTER: Susan remembers attending early LTER Data Manager meetings at the University of Illinois at Urbana-Champaign (1982), Oregon State University (1983) and the Baruch Institute, University of South Carolina (1984). Susan felt a kindred spirit with Walt Conley (one of the JRN lead scientists), Carl Bowser (NTL) and Bill Michener (North Inlet LTER site's Data Manager), and informally, Bill and Susan assumed roles as co-chairs of these early meetings. There were no real data directives at that time other than to insure that LTER data didn't suffer the same fate as IBP data. Data managers realized early on that they would be far more successful advocating for "standards" in data collection and documentation than imposing strict "standardization" as not all PIs embraced the idea of data curation and archival. There was a realization that "standardization" would be best accomplished through "standards" of quality, best field practices and good documentation, and not by defining a single way to collect and process data. Sites were concerned about excellence in their site science and were collecting data in each of the six core areas, but were collecting data in their own way. Telling sites that they had to collect data in one specific way would have been "pushing water uphill". There was a fine balance in getting people to accept that beyond site science there was an added expectation that the sites would need to work together as a "network." Just being a site with phenomenal science was not enough. There was a certain mindset needed in those early days to make a site flourish. The sites that fostered a spirit of openness and demonstrated a collaborative nature were better positioned to meet this added expectation and those characteristics helped define the early success of LTER sites.

The NSF had vowed to prevent the same problem issues with data curation and archiving that had occurred in the IBP from recurring in LTER. Data management failures of the IBP may have been the best friend of LTER data management. Early LTER data management efforts were watched closely by NSF. That is, NSF indicated they knew the importance of data curation and would provide resources and/or guidance where necessary to make this happen. For example, as a means of being on the forefront of GIS technology, the late Tom Callahan (NSF) held a meeting in Washington, DC in the later 1980s to assure that potential software and hardware needs to accommodate GIS were discussed and well-vetted.

Susan believes the solidarity and blossoming of LTER data management was due largely to the annual meetings. She remembers writing a proposal with Bill Michener, early in the history of LTER, which provided funds for the annual meeting of the data managers and which have now become institutionalized.These yearly meetings were critically important because, unfortunately, in the early days of LTER, some of the Data Managers were not recognized as an important member of their respective research teams. The role of the data manager was sometimes marginalized at their own site. The annual meetings helped address this by creating both a sense of community within the data managers (later known as Information Managers) as well as a community of talented and dedicated individuals who were all concerned about the same things – the plots were different but the themes were the same. Focusing on this thematic overlay allowed the data managers to become the poster child for how the network fostered collaboration. The data managers represented a critical mass of talented people with diverse skills, including PhD scientists, professionals and field technicians, which made this community work. This organization of data managers gave a legitimacy to work that we knew was important, but had largely heretofore been seen as merely custodial. The early mini-sabbaticals at LNO were also useful in increasing the visibility of and adding legitimacy to information management work. Susan is gratified that today, data management is taken very seriously and plays a very significant role within the LTER Network and all NSF proposals.

The collegiality of the group was also a key part of its success. Susan remembers a meeting in the mid-1990s where a new participant in the group became aggressively argumentative and took Susan on because he felt he knew better about an issue and Eda Melendez (LUQ) "practically grew out of her skin to exclaim to the participant 'Let me explain to you how we work…'. Everybody became advocates for each other and it was authentic". People shared how similar issues were resolved at other sites and IMs became aware of the issues and were given tools to address them. The meetings were a safe environment where folks could "let their hair down" and ask for help without worry of embarassment. Susan remembers John Porter (VCR) in 1987 introducing himself at his very first meeting and proclaiming 'I'm proud to announce that we have one megabyte of data!' In a meeting at Archbold Biological Field Station in 1996, Susan remembers Jim Gosz (LTER Chair) warning the data managers saying 'the one thing the IMs need to be aware of – you're all very nice to each other and you may even be too polite to each other…'. Even Susan's parents upon hearing mention of 'LTER' (until their dying day) would remember 'those are the nice people that looked out after you'. (This was a reference from Orlando 2002 where John Porter and a group of IMs stayed in the lobby with Susan after the meeting so she wouldn't be waiting alone.)

In 1992, Susan along with Bill Michener, James Brunt, and Tom Kirchner (SGS then CPER) traveled to Beijing, China (recall that this was shortly after the Tiananmen Square protest in 1989) to work with members of the Chinese Ecological Research Network (CERN) to help them organize their information management network.

Bill Michener, James Brunt and Susan teamed up again in 1994 in organizing and producing a Taylor and Francis volume for an international symposium in Albuquerque, "Environmental Information Management and Analysis: Ecosystem to Global Scales", and with John Porter in 1998 for the DIMES conference in Albuquerque, "Data and information management in the ecological sciences: a resource guide". Susan was a coauthor with Bill, James, John Helly, and Tom Kirchner on the seminal Ecological Applications paper,"Nongeospatial metadata for the ecological sciences" in 1997.

Susan also remembers working collaboratively with James Brunt and John Porter at the SEV to help design and participate in the original training sessions for our counterpart IMs from the ILTER, setting the stage for the very successful eco-informatics training sessions we have today. These courses have been offered globally and very ably taught by Kristin Vanderbilt, John Porter, Wade Sheldon, Don Henshaw, Peter McCartney and many, many other LTER IMs. It is interesting to note that Peter McCartney is now a Program Officer at NSF directing, among other things, the bio-informatics program. LTER Information Management has indeed been a successful launching pad for many of our colleagues!

Susan developed a unique professional identity through her leadership of the LTER IM Committee, and this experience likely led to her later posts. Susan spent one year at NSF in 1994 as Director of the then Division of Biological Instrumentation and Resources (now known as the Division of Biological Infrastructure) and chaired the LTER IMC until Barbara Benson was elected chair in 2003. Susan ran very punctual meetings and was amazed how exhausting it was simply to listen carefully! Susan honed her skills as the Committee Chair and could recognize where people's strengths were and led them into positions where they would succeed. Susan's polite and respectful manner was helpful and enabled people to work effectively together. The professional relationships that grew from her work and long-time association with the LTER IM community are some of the strongest, most professional and most enduring that Susan enjoys to this day!

Beyond LTER

In 1998 Susan left OSU to be Department Head of Forest Sciences at Colorado State University. Susan worked with the Shortgrass Steppe LTER site during this period and greatly enjoyed working with Nicole Kaplan. In 2002 Susan left for the Twin Cities of Minnesota to be the Dean of the College of Natural Resources at the University of Minnesota (UM) and continues at Minnesota as a faculty member.

Susan remains active as:

- a member of the LTER National Advisory Board
- past-president and current Board member of the American Institute of Biological Sciences, Chair of the Leadership in Biology initiative and member of the Long Range Planning Committee
- a member of the NEON Board and Chair of the NEON Communication Committee
- collaborating scientist with Judy Cushing (The Evergreen State College) on an NSF–funded interdisciplinary VISTAS visualization project

Susan lives in Veneta, Oregon, with her husband Buster Davis and has a son and two grandchildren.

Sensor and sensor data management best practices released

edit

Corinna Gries (NTL), Don Henshaw (AND), Renee F. Brown (SEV), Richard Cary (CWT), Jason Downing (BNZ**), Christopher Jones (NCEAS), Adam Kennedy (AND), Christine Laney (JRN), Mary Martin (HBR), Jennifer Morse (NWT), John Porter (VCR), Jordan Read (USGS), Andrew Rettig (University of Cincinnati), Wade Sheldon (GCE), Scotty Strachan (University of Nevada), Branko Zdravkovic (University of Saskatchewan)**

Rapid advances and decreasing costs in sensor technology, wireless communication, data processing speed, and data storage capacity have enabled widespread deployment of automated environmental sensing systems. Basic environmental processes can be monitored continuously in habitats ranging from very remote to urban providing information in unprecedented temporal and spatial resolution. Although research questions that may be answered based on these data are very diverse (Porter et al. 2009), the design process, establishment and maintenance of most environmental sensor systems, and resulting data handling have many commonalities.

Realizing that sensor networks are becoming ubiquitous in ecological research and, hence, a new set of skills, approaches, applications, and technologies are required, a workshop was organized jointly by researchers from the Northeastern Ecosystem Research Cooperative (NERC) and LTER information managers in 2011 at Hubbard Brook Experimental Forest with participants from many projects currently implementing sensor networks. An earlier publication reported on the need for an online resource guide as identified during that workshop (Henshaw et al., 2012). The publication by Campbell et al. (2013) on streaming data quality was a major product of that workshop, as was the basic outline for a best practices guide. To date, four sensor training workshops have followed, focusing on remote data acquisition as well as strategies for managing sensor networks and software tools for handling streaming data. The Remote Data Acquisition (RDA) sensor training workshops, co-sponsored by LTER and the UNM Sevilleta Field Station, focused on the field aspect of environmental sensor networks, included hands-on training in basic electronics, photovoltaics, wireless telemetry networks, and datalogger programming. Another RDA workshop will be offered in January 2015, for which more information can be found here http://sevfs.unm.edu/workshops/rda2015.html. These workshops were nicely supplemented by two training workshops that focused on managing streaming data. These were co-sponsored by LTER, NCEAS, DataONE, UCSD, and SDSC and involved trainers from projects as far ranging as LTER, the Kepler Workflow System, the GCE Data Toolbox for Matlab, CUAHSI, DataTurbine and introductions to other open source tools such a the R Statistics Package and the approach NEON is taking. Extensive training materials for these workshops, including power point presentations, example data, and manuals were developed, and some presentations recorded live, all of which can be accessed here http://wiki.esipfed.org/index.php/Workshop_Materials (Henshaw and Gries, 2013).

A working group of practitioners experienced in the entire life cycle of streaming sensor data (sensor network establishment, remote data acquisition, data storage, quality control and assurance, and data access) has met regularly in person and via online forum to initiate the 'Wiki Process' for assembling the Best Practices document. The guide builds on the collective experience of working group members as well as the earlier workshops described above. The working group has released initial documents on the Earth Science Information Partners (ESIP) Federation wiki page (http://wiki.esipfed.org/index.php/EnviroSensing_Cluster).

In its current version, this document on best practices for sensor networks and sensor data management provides information for establishing and managing a fixed environmental sensor network for on- or near-surface point measurements with the purpose of long-term or "permanent" environmental data acquisition. It does not cover remotely sensed data (satellite imagery, aerial photography, etc.), although a few marginal cases where this distinction is not entirely clear are discussed, e.g., phenology and animal behavior webcams. The best practices covered in this document may not all apply to temporary or transitory sensing efforts such as distributed "citizen science" initiatives, which do not focus on building infrastructure. Furthermore, it is assumed that the scientific goals for establishing a sensor network are thought out and discussed with all members of the team responsible for establishing and maintaining the sensor network, i.e., appropriateness of certain sensors or installations to answer specific questions are not discussed. Information is provided here for various stages of establishing and maintaining an environmental sensor network: planning a completely new system, upgrading an existing system, improving streaming data management, and archiving the quality controlled data.

The chapters contained in this guide are structured to provide a general overview of the specific subject, an introduction to methods used, and a list of best practice recommendations based on the previous discussions. Case studies provide specific examples of implementations at certain sites.

- Sensor Site and Platform Selection considers environmental issues, site accessibility, system specifications, site layout, and common points of failure.
- Sensor Data Acquisition outlines considerations and methods for automating real-time acquisition of environmental sensor data from remote locations.
- Sensor Management Tracking and Documentation outlines the importance of communication between field and data management personnel as field events may alter the data streams and need to be documented.
- Sensor Data Management Middleware discusses software features for managing streaming sensor data.
- Sensor Data Quality discusses different preventative approaches to minimize data inaccuracies and quality control and data management practices to identify and properly document problematic data and data quality level.
- Sensor Data Archiving introduces different approaches and repositories for archiving and publishing data sets of sensor data.

As mentioned above, this is a living document, an open source, community supported resource that implements the 'Wiki Process' and everybody is invited to contribute knowledge and experience, provide updates and corrections, or start a completely new chapter with currently missing information. Anyone can register an account with the ESIP wiki and upon approval, may edit existing content. The 'Wiki Process' for amassing knowledge in an organized fashion is well documented for the Wikipedia (see links below). Documentation and guidelines are provided there for reaching consensus through editing or through discussion. Each subject area may be discussed in the Wiki on the 'discussion' tab, or on the ESIP EnviroSensing Cluster mailing list. Overall, most of the '10 simple rules of Wiki editing' apply here as well. We would like to particularly encourage contributions that describe existing local systems (i.e., 'case studies'). Personal experiences and evaluation of products, sensors, software, etc. have not been included yet, but are considered valuable and should be voiced in this forum to the degree that they might help others. A glossary of terms would be useful as well as a list of pertinent publications. Other areas currently not well covered are implementations of OGC Sensor Web standards, aspects of citizens involvement in sensor applications, and cutting edge developments that we are not aware of. However, we hope this effort will provide a forum for lively discussion of the latest developments in sensor technology and that you will find the existing information useful enough to share your knowledge in return.

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.

CUAHSI - Consortium of Universities for the Advancement of Hydrologic Sciences http://cuahsi.org/

DataONE http://www.dataone.org/

DataTurbine http://www.dataturbine.org/

GCE Matlab Data Toolbox by Wade Sheldon https://gce-lter.marsci.uga.edu/public/im/tools/data_toolbox.htm

Henshaw, D., C. Gries, R. Brown, J. Downing, 2012. SensorNIS: Community engagement to build a resource guide for managing sensor networks and data. DataBits Fall 2012. http://databits.lternet.edu/fall-2012/sensornis-community-engagement-build-resource-guide-managing-sensor-networks-and-data

Henshaw, D., C. Gries, 2013. Sensor Networks Training Conducted at LNO. DataBits Spring 2013. http://databits.lternet.edu/spring-2013/sensor-networks-training-conducted-lno

Kepler Project https://kepler-project.org/

LTER - Long-Term Ecological Research Network http://lternet.edu/

OGC - Open Geospatial Consortium http://www.opengeospatial.org/ specifically the domain 'Sensor Web': http://www.opengeospatial.org/domain/swe

Porter, J.H. , E. Nagy, T.K. Kratz, P. Hanson, S.L. Collins, P. Arzberger, 2009. New eyes on the world: advanced sensors for ecology. BioScience 59 (5), 385-397. doi: 10.1525/bio.2013.63.7.10

NCEAS - National Center for Ecological Analysis and Synthesis http://www.nceas.ucsb.edu/

NEON - National Ecological Observatory Network http://www.neoninc.org/

SDSC - San Diego Super Computer Center https://www.sdsc.edu/

UCSD - University of California San Diego, specifically the California Institute for Telecommunications and Information http://www.calit2.net/

UNM - University of New Mexico, specifically the Sevilleta Field Station http://sevfs.unm.edu/

Wikipedia:Ten Simple Rules for Editing Wikipedia, http://en.wikipedia.org/wiki/Wikipedia:Ten_Simple_Rules_for_Editing_Wikipedia

Wikipedia:Consensus, http://en.wikipedia.org/wiki/Wikipedia:Consensus

Wikipedia:List of policies and guidelines, http://en.wikipedia.org/wiki/Wikipedia:List_of_policies_and_guidelines

Sustainable development

edit

Inigo San Gil (MCM,LNO)

Abstract

In this article, we cover some sustainable code theory and practices. To exemplify these practices, we describe the Drupal Ecological Information Management System (DEIMS) working group practices adopted over the course of the last year, detailing the advantages and challenges.

Outline

After a brief introduction and motivation (why would anyone invest in sustainable code development practices), this article shows specific experiences based on sustainable code practices, including a detailed list of sustainable practices adopted by DEIMS. Having lived with these new methodologies for over a year now, we will look back and offer some reflections on what works and what pitfalls we have eliminated from our development pipelines. We emphasize the outcomes that helped our day to day development routine. In the spirit of transparency and full disclosure, we balance the article by noting the overhead and costs that are involved in adopting these sustainable practices. For the impatient, we advance a summary: overall, adopting some sensible sustainable development practices is an investment that you want to make, whether your daily work goes in tandem with an open source thriving community, or whether you work on a small team in custom-code closed source solutions.

Motivation: The good in sustainability.

Why would anyone change his/her project development *modus operandi*[1] and adopt costly practices ? When your project is aimed at a large user base, sustainability becomes critical. You will have to look for efficiencies that will make the lifecycle of your project attain its original goals with the least effort. Aspects to take into account when looking for better sustainability practices include, licensing, development process, project management, community process and marketing. One of the advantages was detailed by Servilla and Costa (2010): openness and the use of a software version control system. Following on openness, perhaps, an important early decision is choosing an open source platform or a proprietary base solution. Some proprietary products may have a clear edge on meeting your project needs. However, your choice may influence the sustainability of the project. Arguments in favor and against open source abound (von Hippel and von Krogh, 2003). At the very least, we should caution that an open source choice is not per-se a winning criteria. Not all open source project are the same: choose a project backed by a strong, active, thriving respectful community. There are many adjectives that we used to constrain 'open source', and each of them is relevant.

Strong. Small open source projects face daunting challenges. Most (but not all) open source projects begin small, and many terminate in the maturing process or even before. Lack of enthusiasm, competition, dwindling support to keep the project alive are common causes of premature dismissal of a promising open source initiative. If you have a choice between adopting a solution sponsored by a strong open source initiative or a nascent weak one, you may want to avoid the risk of the weaker one, being most other parameters more or less balanced. For example, Drupal has over a million registered users with 33,513 developers[2] whereas our earlier attempt (Aguilar et al., 2010) to meet the metadata editor challenge, based on the promising *XForms* technology, is de-facto obsolete with the announcement of discontinued support by Mozilla (2014) and the World Wide Web consortium.

Active. What level of activity? The total level of activity may offer you little actionable information. The activity level has to be viewed in the perspective of the community that supports it, and the balance between users and developers. It is easy to forget about the size of project when gaging the activity level. Large projects will likely have more activity than a small project, but may not necessarily be more active than a small but active project. Measuring activity and activity trends is elusive. Some guides we used: Ask how active and engaged is the community around the components of the solution being evaluated. Grade the overall activity relative to the smaller parts that compose the project. If the information is available to you, evaluate also the context: Examine similar projects of similar size and age. Other factor that may show the level of activity is the release schedule, and the typical content of such releases. Do not confuse activity with stability, stable projects tend to have longer release schedules and fewer volume of changes. For example, all Drupal projects offer clear measures of use, adoption, release and commits. For example, at any given Drupal project page, you will see the following data under the "project information" section: A maintenance status, development status, reported installs, number of downloads, date when the project was last updated and other statistics compiled in a graphs, regarding the maintenance state, such number of new issues, the response rate, the response time lag to an issue, number of open bugs and participants. Per se, these statistics do not reveal much, but in the context of other similar projects, the statistics are of enormous guidance. Not all projects offer tools to gage the health of the system, but the mere existence of such tools should weigh in your decision.

Thriving. Is the community around this product a happy, growing community, or did it stall? Is it plaged by disenchanted participants? Are there many forks that surfaced owned by disgruntled participants? In contrast, a thriving community moves forward and the members irradiate enthusiasm about their involvement.

Beware of open source projects whose major contributions come from developers on the payroll of the project. Lack of substantial voluntary contributions may denote a lack of faith and passion for the project, ingredients that increase the long-term viability and sustainability of any software project.

Respectful. Good projects take in feedback and adapt to necessary changes. *Holier than thou* attitudes and a tight grip on the project code denotes dismissive and disrespectful attitudes towards the constituents, and often a sign that the project lacks any future. Communities built on a code of conduct that fosters integration, diversity and participation are more likely to thrive that communities controlled by a few persons that leave little room for innovation and progress. The diversity of the community is also a good indicator of the community health, specially the gender ratio, as Petersen (2014) argues in her analysis of the influence of diversity on open source communities in the US and abroad.

Beyond the choice of open source project as a base framework, there are several choices you can make to ensure viability of your project. In addition to the aforementioned version control system, you may want to ensure your project is ready to be deployed. No custom scripts, no obscure licenses (but procure a copy of a known license): streamlining deployment is key in the adoption process. A complex installation may erode the confidence of the adopter. One of DEIMS key objectives was to package the product in the simplest form possible. The DEIMS adopter has the choice of building from source, or to deploy a pre-built package. Long-winded documentation may detract the adopter. If the install process resembles what the adopter may expect, your project's chances of pleasing the adopters increase. For example, installing the DEIMS package is almost the same as installing Drupal, however, this was not the case in the first DEIMS release.

Sustainable practices include the creation of quality documentation in several formats. Good documentation that targets developers and users encourages them to commit to the use and maintenance of the software. Documentation is not just a set of guidelines, it is a reassuring sign the project meets one of the critical pilars of sustainability. Ironically, documentation is one of the first things dropped from the task list when pressed by tight schedules. Software project documentation is akin to the metadata of a dataset. Make sure you budget resources to produce some documentation in the form of how-tos, videos, guided tours and the likes.

As an added bonus, the longevity of your project may be helped by the addition of product-oriented training services to your adopters. Hands-on, relevant exercises offer the attendee a chance to actually practice. If possible, design the exercises to cover issues of interest to your attendees to make a more satisfactory user experience.

Another component of sustainability within the development process is a public repository. GitHub was deployed in 2008, and today it is billed as the largest code-project host in the world, with over 13 million at the time of writing (Wikipedia, 2014). GitHub success is based on the ease to create, use and share code. Much more so than another long standing repository known as Sourceforge. In addition, attributing credit and tools for cooperation are well implemented. An excellent collection of indexed help-pages would help the newbie to start using GitHub in no time. As a plus, GitHub lowered the barrier of git-knowledge requisite. Simple apps can be downloaded for any platform. The GitHub apps transform the relatively complex git operations with usable and intuitive workflows. Giving credit to contributors in GitHub is easy which is both stimulating and challenging for the developer. GitHub places the details of the development process the forefront of the project, providing a clean view of the chronological, developer-centric activity. *Branches, commits, pull requests* are all acompanied by good documentation for the visitor.[3]

Embracing sustainability: from theory to practice

In this section, you will find a summary of what DEIMS did differently during the last year in terms of sustainability. We highlight practices since the inception of the project, and how those have been changed over time.

A unified LTER Information Management system. Perhaps the most important sustainability aspect that DEIMS continues to exercise is the development of a common information management base platform for LTER sites (Gries et al., 2010). There is nothing of more value in my mind than to keep a united team in the face of change and challenge. Not as a traditional union, but as a cost-savings effort. The DEIMS team acted on the realization that the LTER sites had been solving the same problems with different approaches and technologies over the last few decades. Decades ago, this decentralized approach made sense:

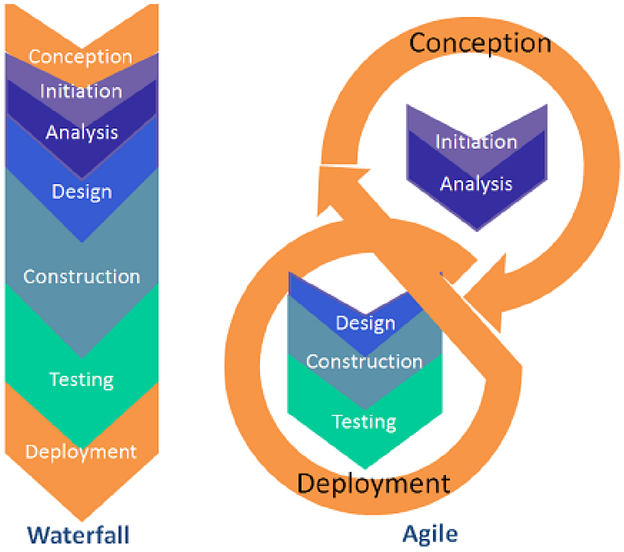
development choices were driven by the site in-house expertise. After all, each LTER site is responsible primarily before the NSF review panels about their own site-specific work.

In the age of decentralized information services (the cloud) and collaboration, access to information and sharing has never been easier. In fact, it is so easy to collaborate that it is hardly justifiable to work alone for a solution of problems for which there are existing solutions. Information sharing barriers are low enough to justify an investment on your LTER network partner products which in term produce a greater Return to the Investment than facing the same challenges alone.

Adopted Agile as the development process while we were intensively scaffolding the new major version. This meant we have daily scrums with the developer team, where LTER participation resulted in important efficiencies for the project. Agile (Highsmith, 2004) principles can be summarized as close cooperation, rapid development and delivery of the highest value product at each development iteration process. The *cooperation* and *rapid* implies that we were intensively engaged with a developer team that moved faster than what could be imagined. If you want to experience the returns that Agile promises, your team will have to get engaged in the agile process actively with self-motivation. How does Agile relate to sustainability? IT projects need to move to the pace of the underlying industry. Waterfall, which was the development process followed in earlier decades, where all milestones of the project were defined at the beginning, turned out to be inflexible in terms of adapting new trends and components, frequently rendering obsolete projects before the delivery date. For example, a five year schedule to release a software product is an anacronism in an industry where we see unexpected changes every six months in major products. Most of the major software corporations changed to Agile which creates more relevant and sustainable projects. DEIMS lacked any structured development workflow in the initial stages. It was largely two Drupal enthusiasts guided by two information managers, using Google Code as the public code repository.

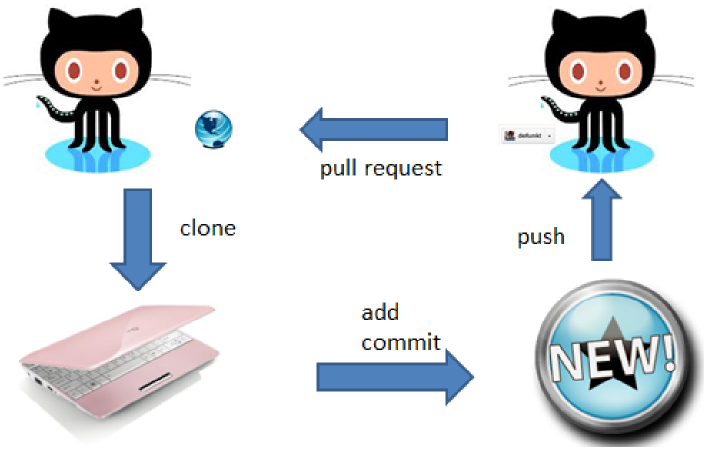
Some argue waterfall has a place for large and long cycle projects, however, there is no clear consensus to our knowledge. We found many special beneficial features in the Agile process - we could test a critical piece of the architecture before any dependency was built around it, saving us time. We could also evaluate the real cost of deploying features before committing entirely to the task completion. This pre-evaluation was used to deck some costly features, such as a responsive template for a slide show.

Fig 1. The following is a schematic that compares two major approaches to software development process. Note that waterfall starts from the top, and does not move into the next stage of development until the previous is complete. In contrast, in Agile the overall project is broken into atomic components, and each component goes into the entire development cycle, repeating the process constantly, and allowing to define and revisit the highest value component at each stage.



Adopted GitHub as the project repository. We have a core project called DEIMS where the common core is installed. LTER sites need certain extensions that sometimes are unique to the living model of the site. All extensions and related customizations (such as the migration code from the previous information management sources) are also hosted in GitHub (see the Arctic LTER, Luquillo LTER, International LTER, the McMurdo LTER, the North Temperate Lakes LTER and Sevilleta LTER). As mentioned earlier, before this release, DEIMS used Google Code as public repository. While Google Code served well to accept community contributions and to market the product, it lacked the awesomeness of GitHub development workflows.

Fig 2. A simplified schematic of a GitHub based typical development workflow.



An eye on the near future to decide the practical implementations for the immediate future, a *condicio sine que non* for any sustainable software effort. The best predictive knowledge you can use is based on science. In the absence of this knowledge, we found useful some near-term knowledge. We enrolled a development team with deep knowledge in the new version of Drupal. In contrast, DEIMS earlier release was did not account for changes. You should design keeping an eye of the advancements of your supporting framework, which in software projects, may be a mix of Java, Javascript, PHP, derivations of such languages and many others components of your final stack. Where is the main trend in the next release? What features are being deprecated, and what tools are being matured?

One Drupal aspect I value the most is the sincerity of the Drupal steering committee. For the last three years, I listened to the "state of Drupal" hour-long talk offered at the main Drupal Conference hosted in the United States. Drupal founder Dries Buytaert (2012, 2013, 2014) analyzes the placement of Drupal in the current web context, and where, as a community, Drupal should be positioned. This analysis ensures that while the Drupal community works intensively on the current platform, the designs for the next major release are being worked on and applied. In the last four years, the Drupal community has experienced two major version changes, the Drupal-seven release has usability as the main focus while the forthcoming Drupal-eight release emphasizes the underlying engine. With D8, Drupal makes a brave leap from a procedural-leaning software base, to a object-oriented code base. DEIMS has been developed with Drupal-seven in its core, but newer critical components were developed with the next iteration in mind, creating a DEIMS future-proof release (Reid, 2014). Some DEIMS components that are object-oriented include the Ecological Metadata Language, the Biological Data Profile, the International Standards Organization 1911X and the harvest list services. DEIMS has implemented seventy-five percent of the list of Drupal-8 module backports that Reid covers.

Steering the human capital. A sustainable project must have good governance. DEIMS mimics well the Drupal diversity within its ranks. DEIMS works with open input from its constituents, stakeholders and public: The issue queues both at GitHub and drupal.org are not limited to an LTER subset of people, but open to the world for feedback and therefore, open to the world for contributions. Questions are addressed openly and accepted in terms of merits, welcoming development from those who conform to the Drupal code practices.

A serene and playful arena. Sustainability relies on the people who make up the ecosystem. Just like in the anthropocene era, in software development the human dimension plays a fundamental role. How we work as a group determines the viability of a project. Scholarly analyses (Koziolek, 2011) often forget to mention this critical pillar of sustainability, perhaps because of the difficulty to objectively quantify and assess the human role. One surprising aspect of the crowded issue queues of Drupal is the relative absence of un-constructive comments. Experts are patient and point the newbies in the right directions, and overall there is a to-the-point focus on the queue. No deviations to the task at hand. As a result, any Google search would likely land you in the relevant topic, sans the spam or the abuse. Part moderation, part code-of-conduct, the forums are inviting and useful. DEIMS added issue reporting to both GitHub and drupal.org, while continuing the Google Groups and email for group communication.

Reuse Readiness. As Marshall (2008) argues "The reusability of software products is a factor often neglected in common measures of technology maturity". When you consider your project, ask yourself: How can a new developer re-use and re-purpose your software project. Can he or she extend your project framework to make it conform to new requirements? DEIMS design stricltly followed this pattern. DEIMS re-uses the contributions of 86 extension to the 41 modular pieces in the Drupal core. The vast majority of the project is re-usability of existing, living and actively maintained code. DEIMS only created 11 modular extensions that are entirely custom, some of them to meet du jour[4] requirements, such as a content export specification or the consumption of external services, such as the client for the LTER unit dictionary.

Exit strategies. A sustainable design must have a clear technological exit strategy. DEIMS has several paths that an adopter can use to amass and export its content. DEIMS can provide the entire cross-related content using SQL queries rendered in comma delimited files. DEIMS can use services, both expressed in XML schema conforming services, schema-less XML and JSON, which is favored by most modern applications. DEIMS also used the Drupal migrate framework, dividing content into sources, destinations and mappings, and implementing plugins to adapt the sources, destinations and even tweak the mappings for the oh-so-common exceptions.

Challenges

Changing habits is not trivial. Think of the many times you go to the grocery store with your eco-friendly bags[5]. Better yet, look around you in the checkout lane: how many people still load their goodies in plastic or paper bags. If your checkout lanes are not much different from those in Albuquerque, you may be wondering why such a simple habit is not quite adopted yet, despite the benefits. The analogy holds for software development. Change brings the deepest fears on us, yet we need and embrace change to succeed in providing society with the best predictive knowledge that science can offer.

Versioning systems such as Subversion or Git have been around for a long time, yet, some of us still manage to skip the learning curve and keep our custom and idiosyncratic personal stash of versions and what-did-i-do-yesterday confusion.

Documentation is one of the first things dropped from the task list when pressed by tight schedules. Yet, a software project without documentation is like a dataset without context (metadata).

Complexity. The data generated by LTER is complex, heterogeneous, incomplete and unwieldy. The systems to document these data and information are not surprisingly complex. With each iteration that makes the use of the information system, the complexity of the management system increases.

Funding model and structure. Information management at LTER as a community is not far from the cooperative model necessary to thrive in this increasingly complex information delivery ecosystem. However, funding allocations present a challenge that has to be overcome. DEIMS managed to bring together funds from several sites to develop a product that benefits all, however, the mechanisms we used to fund the working group and project are not the best, nor easy to implement.

Summary and Concluding Remarks

We conclude the short research and experiences on sustainability narrated here with the hope to open a dialogue on the important sustainable development topic. We saved many hours by following practices that required us to learn something new (Agile, Gitflow, GitHub, the Drupal API, to mention a few). A final few take home ideas will be echoed in the next bullets.

Open your code to the wider community. Let strangers tell you about bugs, problems and possible enhancements. Let people request a new feature, suggest changes, and also, let them contribute fixes to your living project.

Do not build from scratch. Whatever your need, ideas, or recipes, all those have been invented before, several times. Adopt a seasoned, open source solution that is synergistic with your project goals. Joining and extending an existing project will give you a heads-up in community, project maintenance, sustainability and maturity from the start.

Evaluate and test your options. When embracing an existing project, make a list of pros and cons. Look for red-flags, predictable impediments that would doom your courageous initiative.

Find what moves you most in the sustainable lifecycle, and invest your efforts in there, most developers find their peak productivity when they do things they consider fun or engaging or challenging.

Finally, a shameless plea for collaboration. Please join in with your excellent expertise and enthusiasm to power the circulation of knowledge for the ecological community. There are countless ways that you can contribute, here is a short list.

Documentation big and small. Issues you find. Features you miss.

Testing: Can you test some aspect that?

Developing: Ticked by the ISO dataset object? The queue that waits for external services? The deploying process? there are ample new (even for Drupal) ways in which we used the Drupal API. If developing is amusing to you, you'll find many opportunities in a code of this magnitude. Check the GitHub issue queue and the Drupal issue queue for DEIMS.

Translations: You could help translating the project (the components, from documentation to labels). Drupal has an awesome translation management system!

Funding: We are opened to your sponsorship.

You are most welcome in joining the DEIMS grass-roots initiative, and we would love your (anticipated) contributions.

Footnotes and Comments

[1] Modus Operandi aka MO. Oh, yes, it is latin for routine, habit, common method, standard protocol or more literally, method of operation.

[2] Number of active Drupal developers as indicated in the project front page. Resource visited on June 2014 at http://drupal.org

[3] One interesting consequence of this level of transparency at GitHub is how open source is changing how we are hired. GitHub and other public repositories provide employers with the ability to judge a developer by the amount and quality of her contributions. The footprint we leave on the web provides a metric which may be arguably more important (see Lopp, 2013) than the old 'list of mastered computing languages' and 'roles' that appear in classic resumes.

[4] From Merriam-Webster: popular, fashionable, or prominent at a particular time <the buzzword du jour>

[5] Bagging in western Europe is optional: Do you want a plastic bag? We got them! ranging from 10 to 20 euro cents per bag and just as conveniently available in the checkout lane. The bagging experience spectrum is substantially different.

References

Aguilar, Raul; Pan, Jerry; Gries, Corinna; San Gil, Inigo and Giri Palamisamy. *A flexible online metadata editing and management system* (2010) Ecological Informatics. 5. 26–31. Resource at: http://caplter.asu.edu/docs/papers/2010/CAPLTER/Aguilar_etal_2010.pdf

Buytaert, Dries. *Denver DrupalCon Keynote*. (2012). Resource at: http://www.youtube.com/watch?v=RddJvlbSY88

Buytaert, Dries. *Portland DrupalCon Keynote.* (2013). Resource at: http://www.youtube.com/watch?v=PCLx4fRHmCk

Buytaert, Dries. *Austin DrupalCon Keynote.* (2014). Resource at: http://austin2014.drupal.org/keynote-dries-buytaert

Callaway, Tom. *How to know your free or open source software project is doomed to fail.* (2009). Resource at: http://spot.livejournal.com/308370.html, last accessed June 2014.

Gries, Corinna; San Gil Inigo; Vanderbilt, Kristin and Garrit, Hap. *Drupal developments in the LTER Network* (2010), Databits, Spring 2010 ed. Resource at http://databits.lternet.edu/spring-2010/drupal-developments-lter-network

Highsmith, Jim. *Agile Project Management: Creating Innovative Products* (2004). Addison-Wesley.

Hippel, Eric von, and Georg von Krogh. *Open source software and the "private-collective" innovation model: Issues for organization science.* (2003) Organization science 14.2 : 209-223.

Koziolek, Heiko. *Sustainability evaluation of software architectures: a systematic review*. (2011) Proceedings of the joint ACM SIGSOFT conference--QoSA and ACM SIGSOFT symposium--ISARCS on Quality of software architectures--QoSA and architecting critical systems--ISARCS.

Lopp, Michael. *The Engineer, the designer and the dictator*. (2013) Portland DrupalCon. Resource at : http://www.youtube.com/watch?v=rK4Om-_My7Q

Marshall, James J., and Downs, Robert. R*euse readiness levels as a measure of software reusability*. (2008). Geoscience and Remote Sensing Symposium, 2008. IGARSS 2008. IEEE International. Vol. 3. IEEE.

Mozilla, *Archive of Obsolete Content* (2014) https://developer.mozilla.org/en-US/docs/Archive/Web/XForms

Penzenstadler, Birgit; Bauer, Veronika; Calero, Coral and Franch, Xavier. *Sustainability in software engineering: A systematic literature review.* (2012). 32-41. 16th International Conference on Evaluation & Assessment in Software Engineering, Ciudad Real, Spain.

Petersen, Erynn. *Austin DrupalCon Keynote*. (2014). Resource at: http://www.youtube.com/watch?v=zoia8WZ6q5w

Tate, Kevin. *Sustainable Software Development: An Agile Perspective* (2005). Addison-Wesley. ISBN:0321286081

Reid, Dave. *Future proof your Drupal 7 site*. (2014). Watch resource at http://www.youtube.com/watch?v=z3_MqGLjqkA

Servilla, Mark and Costa, Duane. *Openness and transparency development lter network information system* (2010). Resource at: http://databits.lternet.edu/spring-2010/openness-and-transparency-development-lter-network-information-system

GitHub, Wikipedia (2014) Resource at http://en.wikipedia.org/wiki/GitHub

Commentary

LTER Information Management– Past, Present and Future – a Rough Timeline

edit

John Porter (VCR)

In considering where LTER Information Management is going, it is valuable to think a bit about where it has been.

When LTER started:

- Only a few investigators had personal computers, most of them had black-and-white, text only, monitors
- No one had electronic mail
- Network connections were all via modem, and ran at 0.0003 to 0.0012 megabits per second
- Floppy disks would not fit in your pocket (5-1/4" or 8" across)
- Visicalc was the dominant spreadsheet program
- A very large hard drive was 20 MB
- No LTER site shared data in any systematic way

When the Internet became available to scientific users
(circa 1989):

- Only a few investigators had connections to the Internet
- Tools to use the Internet were primarily Telnet and FTP
- There were no graphical user interfaces for working on the Internet
- The first LTER-wide personnel and bibliographic databases started to be assembled
- Many email systems could not intercommunicate with one another (e.g., Forest service could not talk to universities) so LNO IM Rudolf Nottrott set up an LTER-wide mail forwarding system that crossed these boundaries, along with the first email lists
- The LTER Network Office was founded
- A "minimum-standard installation" (MSI) was established focusing on getting GIS capabilities at every LTER site. It was supported by a series of large NSF technology supplements.
- The LTER Network was working to create its very first Data Catalog, which was limited to 10 datasets per site
- The LTER Network prepared the first guidelines for site data access policies
- Only one LTER site (HBR) had a "Bulletin Board" system for sharing data via dial-up connections
- LTER Databits was started
- Work on establishing a minimum standard content for LTER metadata started

Verbatim
DataLife
Floppy Disk 5¼" – 0.36 to 1.2 MB
Cassette Tape (1989)
Floppy Disk 3.5" – 0.72 to 1.38 MB

The "modern," Internet-centric version of LTER came into being in 1994 with the advent of the first web browser. At that time:

- Many sites were operating online systems using Gopher and later HTTPD servers
- The LTER Coordinating Committee mandated that each site share at least one dataset online
- Many LTER sites had developed data access policies governing how data from the site could be shared
- LTER Information Managers had defined a list of common metadata elements
- ClimDB started to be planned (1996)
- The idea of a network-wide information system was adopted by LTER
- LTER Databits went on a 5-year haitus, returning in 1999
- Hard drives are typically less than 100 MB in size

The age of network data systems arrived in 2002:

- The Ecological Metadata Language standard version 2.0 was released
- The first versions of the Metacat data catalog and Morpho EML editor were produced
- The previous year, the LTER Lead Principal Investigators had added "sharing data" to the list of LTER Network goals
- The line between web page and database was beginning to be blurred, content management systems began to be seen
- Hard drives are approaching 1 GB in size

More recently (2010-2014):

- The LTER Data Portal moved to use the PASTA system, assuring better quality metadata and data and more reliable downloads
- LTER increasingly uses web services to share information among programs running at many different sites
- The Drupal content management system is used by many sites, and specialized tools are developed for use at ecological research sites
- LTER Information Managers have established controlled vocabularies for data keywords and units
- LTER data is included in DataONE
- Multi-terabyte hard drives are common
- Many LTER systems run on virtual computers

The next decade:

- Data collection is increasingly digital from end-to-end, with automated sensors and electronic field notebooks coming to dominate paper forms
- Control of field tablets and computers will be via audio or gestures
- More LTER sites will start to use a variety of drone aircraft for real-time remote sensing and for *in situ* measurements using drone-deployed sensors
- Institutions start to evaluate research productivity using citations of published datasets, as well as traditional manuscripts
- Now that every researcher needs to have a data management plan, more commercial products and services aimed at meeting the needs of scientific researchers will be created
- Researchers will start to use standard ecological datasets, such as those produced by NEON

The next century:

- Few LTER researchers will have anything recognizable as a computer of today, but will instead have a variety of interface devices that link them to cloud-based resources
- Computer-mediated semantic interpretation will help to automate assembly of data for analysis
- The data point (with associated metadata), rather than the dataset, will become the focus of information management efforts. A query, rather than selecting datasets that qualify under search conditions, will return a collection of individual data points assembled from a large number of individual studies
- The analysis and interpretation of large amounts of data will be the primary challenge, not the collection of data

Don Henshaw made several helpful comments on this timeline, but any errors remain my own.

Reading the bones: a few thoughts about the future of research information management

edit

James Brunt (LNO)

"Where the hell's my flying car?" - Steve Earle - from 21st Century Blues"

I'm the first to admit that I'm not a great seer. I missed the whole smartphone thing and I still don't understand social media. Sure, I predicted the digital data deluge but who didn't? I also predicted that we'd have analytical and visualization tools beyond our wildest dreams. Whoops. My predictions when right though have been less prophetic and more following trends out to some logical future with a judicious application of Moore's Law.

If I had to make just one prophetic statement it would be that **site information management (IM) will get more complex.** If the trends I watch were linear, Information management should getting more automated, be focusing more on workflow process and less on site-based IT infrastructure, and be relying more on community resources. I say this because we are going to have cheap accessible cloud-based computing and storage infrastructure, more community resources, like PASTA, to draw from, more tools available to help reduce the effort spent on tasks that are repetitive and common across sites, and more community standards available for adoption. These trends all point towards increasing IM functionality and lessening site IM burdens. So, you ask, doesn't that mean that site IM should be getting simpler? Not at all, I predict that Glass' Law[1] will prevail and rather than simplifying IM by taking away, we make site IM more complex.

I offer these additional predictions based on trends (and hunches) that will lend some credibility to the statement above:

Budgets will become tighter. Ouch, I know. But as more of the broader research community implements data management plans the overall cost of data management to the funders will go up. Unable to sustain the introduction of data management practice where there previously was none the funding agencies will be looking for economies of scale. This will drive the availability of community resources and centralized tools. The initial use of these tools will be driven more by budgets than popularity. Site management teams will have to weigh the cost and benefit of doing more centrally.

Site information management will rely more heavily on local software skills. In the endless effort to bridge the app gap it's still easier to tool up an application for some site-based data need than to search around for and adapt an existing tool. True or False? I believe true for the most part because of the many site specific constraints inherent in most systems. If finding and adapting existing tools was easier this might be less true. Look for more job announcements with an emphasis on multi-faceted programming and software integration skills.

Data quality and data assurance will become more important. Consumers of data products are always looking for data of known quality. Defined data quality metrics will become more common and their use will inform both producers and consumers of data. Richness of metadata and institutionalized checks will be major factors in predicting data quality. Site information managers will continue to be the frontline defense of data quality.

Open data will pervade and prevail. Don't say you didn't see this one coming. Government, funding agencies, reviewers, colleagues, and the public will cease to tolerate obfuscation through endless registration, policy, and paywall pages between them and data.

Data products will become more important. As a result of open data there will be a greater emphasis put on synthesized data products and reduced scrutiny of raw data publication. As the tsunami of open data crashes upon the rocks there will be outcries and pleading for value-added products that summarize and give meaning to all the raw measurements. Site information managers will be called upon repeatedly to contribute to and support the process of developing these products.

There will be more standards - there will be more innovation. This may seem like a non-sequitur to the IM mind but consider my proposition that useful standards are the ultimate resting place of surviving innovation and not the death of innovation as can be the case with standards that are not empirical in design. Good standards emerge as we use standard approaches to processes that have been originally established as innovation and have survived by popular adoption. This process makes room for more innovation and we know that ecology is a discipline that thrives on innovation. Site IM is central to both the adoption of standards and the implementation of innovation.

Inset Box 1 - What's trending and not for the future of IM?

What's trending?	**What's not trending?**
Data quality	Website quality
Data products	Raw data
Open data	Paywalls
Software skills	One-size fits all
Standards ("Good")	Standards ("Bad")

Footnotes

[1] Robert L. Glass in "Facts and Fallacies of Software Engineering"- For every 25 percent increase in functionality of a system, there is a 4X increase in the complexity of that system.

The Future of Archiving (Research) Data

edit

Ryan Raub (CAP)

The biggest concern for long term archiving data is preservation, how to continually store more and more. With an ever growing volume of data, we need to adapt storage systems that can grow proportionally as well. This is when data collections start to span multiple machines and the existing methods for distributing resources create an unnecessary increasing management overhead.

There is a culture of data storage systems that have shifted away from managing single sources and towards more organized distributed methods for file systems. Distributed models are inherently more complicated, but we can handle this by using abstraction layers that hide some of this complexity and in turn give us more powerful tools to work with. These systems can also give you some big advantages in terms of durability, performance, and capacity; which are all very desirable for data preservation.

Lets talk about some of the abstractions that get introduced, the first is a basic one: How to identify a file? On your personal computer you can uniquely identify a file by it's name and it's location. With a distributed system, that same file exists on multiple computers in different places. What we can do instead, is identify a file by its contents rather than where it is. It is a little counter intuitive to identify to a file by it's contents because of its size, but we address this by using it's contents to compute a (fixed size) hash[1] as a derived unique representation of it. Once we make this transition to working with files based on their hash values we can now identify and talk about a file and not need to know it's actual location(s).

Now that you have this hash identifier for a file: How would you ask for the file itself? Popular existing methods of file transfer rely on a URI[2], which we can and should still use. But we can do better; instead we'll want to make a request to a "matching service" (which runs on all of the nodes of the data network) that can tell us which computers have this file. There will be several sources for every file (redundancy is a requirement) and we can leverage the internal redundancy of this system to aid performance by asking each source for a different part of the file at the same time. Once you collect all the pieces you can recombine them and verify that the end result is what you requested by its hash value.

Adding capacity to this system is as simple as adding new nodes[3] or expanding allocations on existing nodes. If we had a need to provide anyone with faster access to this data network, all we would need to do is create new nodes near their network and optionally set priorities to have it replicate relevant data for them.

This underlying distributed file system still needs to rely on search engines for discoverability; how else are users going to know that a file exists, let alone which file they want. This is an interesting and similarly complicated part of the system that is still being refined and developed today.

I would also like to comment on other popular reference systems; we're currently in the golden age of DOI[4]s. These provide a great authoritative source and a matchmaking resolution service that can be maintained. However they do not provide any way to prove that where they directed you is what the author is actually trying to reference. With just a few more characters they could cite a hash value of the referenced file, uniquely identifying it and providing a method of verification. Another desirable by-product of technique is versioning becomes a non issue; a particular files hash will always only identify that same file, any change regardless of how small will have a different unique hash.

As an example of how this system could be applied to create an data network for the LTER: Each site would provide an node(s) to host their own sites data (given higher priority) and some amount of data from other sites (parity). As a whole the network would benefit from having many geographically separated resilient copies of every version of each file managed automatically. The network improves its disaster recovery potential as well as its overall performance, availability, and storage capacity.

These systems do exist currently and their adoption in the research community is slow. Looking forward with increasing demands and growing amounts of research data, the adoption of these types of systems is now, before we have serious issues. There are many more aspects of data archival that I don't have room to address in this article, such as; file formats, standards, and compatibilities. If you are interested in this topic and want to participate, feel free to contact me.

Footnotes:

[1] - More about hashing: http://en.wikipedia.org/wiki/Secure_Hash_Algorithm

[2] - Uniform Resource Identifier: http://en.wikipedia.org/wiki/Uniform_resource_identifier

[3] - Node: A single computer that contributes computational, storage, and network resources by participating in a larger network of nodes.

[4] - Digital Object Identifier: http://www.doi.org/

Good Tools And Programs

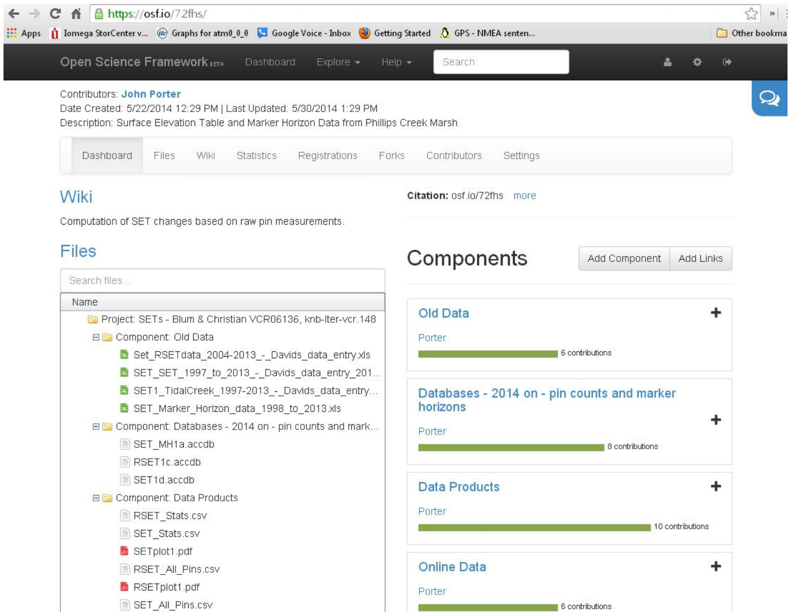
The Open Science Framework

edit

John Porter (VCR)

One of the challenges in any complex process is keeping all the steps straight and keeping track of the files, notes, email etc. associated with a project. The Open Science Framework (http://OSF.io) is a new (beta) web-based tool for helping to manage the scientific process, and to share data. It is being developed by the newly started Center for Open Science (http://cos.io) to help promote reproducible scientific results and sharing of data.

The Open Science Framework provides an easy-to-use and attractive web interface for managing the scientific process - from the formulation of hypotheses, to the collection of data and its analysis. How does it work? Users can define "projects" each of which have one or more "components" (e.g., hypotheses, methods, procedures, data, analyses, and communications). Each component contains a WIKI and a place to store relevant files. Components can be nested or shared with other project contributors.



Contributors: John Porter
Date Created: 5/22/2014 12:29 PM | Last Updated: 5/30/2014 1:29 PM
Description: Surface Elevation Table and Marker Horizon Data from Phillips Creek Marsh

Wiki

Computation of SET changes based on raw pin measurements.

Citation: osf.io/72fhs more

Files

- Project: SETs - Blum & Christian VCR06136, knb-lter-vcr.148
  - Component: Old Data
    - Set_RSETdata_2004-2013_-_Davids_data_entry.xls
    - SET_SET_1997_to_2013_-_Davids_data_entry_201...
    - SET1_TidalCreek_1997-2013_-_Davids_data_entry...
    - SET_Marker_Horizon_data_1998_to_2013.xls
  - Component: Databases - 2014 on - pin counts and mark...
    - SET_MH1a.accdb
    - RSET1c.accdb
    - SET1d.accdb
  - Component: Data Products
    - RSET_Stats.csv
    - SET_Stats.csv
    - SETplot1.pdf
    - RSET_All_Pins.csv
    - RSETplot1.pdf
    - SET_All_Pins.csv

Components

Old Data
Porter
6 contributions

Databases - 2014 on - pin counts and marker horizons
Porter
8 contributions

Data Products
Porter
10 contributions

Online Data
Porter
6 contributions

The attractive interface makes it relatively easy to maintain, good, real-time notes, sets of documents and files. But there is much more under the hood. Each activity by a contributor is logged, so it is easy to see who is working on what. Additionally, files are versioned, so that it is possible to go back and look at earlier versions of programs or data files. Providing an auditable project makes it easier to create a highly creditable scientific project. When the project is completed, it can be made public, sharing the entire scientific process with the community.

One possible challenge is the task of keeping all the files on the web-site up-to-date. However, OSF also works with Dropbox, FigShare and GitHub, so that files in an OSF project are automatically synced. The only downside is that such changes are not currently logged.

I've been experimenting with OSF for managing complex dataset workflows - where data collection, curation, QA/QC and analysis are all done by different people, often in different places. The hierarchical structure is flexible, allowing components to be added as needed, but what I especially like is the ability to fully annotate components and files. It makes it much easier to keep abreast of progress on a project when I can't work on it every day. It also is better than just sharing Dropbox with a collaborator, because the WIKI component allows additional opportunities for annotation.

Managing Bibliographic Citations for LTER

edit

John Porter (VCR)

Publications are the lifeblood of science and academia in general. For this reason, keeping track of publications related to an LTER site is critical. The Virginia Coast Reserve (VCR) LTER has for some time used an aging version of EndNote. But we were curious if there was something better. So we sent out a query to LTER Information Managers asking what they were using.

First, we defined what we thought a “perfect” system would look like:

1. Ingestion
   - automated citation ingestion given a DOI
   - easy to import citations from other sources (e.g. web of science, Google Scholar)
   - easy to cut-and-paste citations emailed to me
2. 2) Display
   - easy to produce, up-to-date attractive web display
   - optional sorting criteria (author, title, year)
   - simple search
   - advanced search (field specific) (author, title, year, keyword)
   - filter by type
   - numbered display (to make counting easy)
   - includes place to link to PDFs of papers
   - includes place to link to associated datasets
   - clickable to display abstracts etc. (if available)
3. Export
   - needs to be able to produce Endnote Export file for importing into the LTERNet Cross-Site Bibliography
   - individual or lists of citations to standard citation exchange formats suitable for input into other systems

Here is a summary of the responses we received:

Corinna Gries at the North Temperate Lakes LTER reports that they also use Endnote for ingestion of bibliographic data, because it exports easily to both Drupal and the LTERNet Cross-Site Bibliography. Drupal provides many of the sorting, searching and querying functions for general users to access.

Margaret O’Brien at Santa Barbara Coastal LTER reports that they use Ecological Markup Language for maintaining citations. The additionalMetadata tag contains management-related information, including related dataset ids, agency reported to, project name/code and report year, and Booleans for completeness, online status and acknowledgments, Stylesheets handle sorting and filtering for the web, and for transformation to an EndNote export format for importing to the LTERNet Cross-Site Bibliography.   Although this is adequate, she is in the process of migrating the bibliographic data to a PostgreSQL database to better facilitate queries for NSF reports and cross-links between people, research projects and datasets. She will keep the citation itself in EML, using the PostgreSQL XML data type, so existing XSL template can still be applied. She would definitely like to see something that could automatically export citations for Report.gov (Missed on the list of “perfect features” above)!

She also noted that it would be useful to have a discussion of various policy issues related to what constitutes a journal paper (i.e., level of peer review), and criteria for what constitutes a publication related to a site (or not).

James Connors at the Palmer and California Current LTER sites reports using EndNote for basic bibliographic management, but that they have a set of Python scripts that parse an EndNote Export file into a custom schema within their relational database. That database is then used to feed their web page. Using a relational database lets them easily link citations with personnel and the data catalog, while maintaining the original data in Endnote facilitates updates of the Cross-Site Bibliography.

Finally, Jason Downing at the Bonanza Creek LTER site uses a custom MySQL schema to manage publications as a part of their “meta-database.” Thus, as for James Connors, allowing bibliographic data to be easily linked to personnel and data packages. They use a ColdFusion service to dynamically format the data for web display. They use simple PHP-based forms to provide a web interface for input of references, with additional cleanups applied directly within the database. They also have scripts for generating EndNote Export formats for participation in the LTERNet Cross-Site Bibliography.

As you can see, there are a wide range of approaches used within the LTER Network. However, there are some commonalities. One is that input to the bibliographic database is primarily through a single portal (entry into a single EndNote database, or directly into EML), indicating that IM or other designated project personnel are responsible for the data entry. Presumably this is because of issues of accuracy and duplication that arise when data come in directly from investigators via web forms. This mirrors our experience at the VCR LTER, where formerly we used online forms to solicit publications from investigators, but soon found that we would get many duplicate, or worst yet, near-duplicate versions of citations accumulating in the database. The time required to clean up the citations was more than the time required to input them consistently and correctly. Another commonality is that bibliographies needed to be transformed into different versions for different uses (e.g., web display, reporting, and inclusion in the LTERNet Cross-Site Bibliography). A variety of tools, from content management systems to
hand-written scripts were used.

At the VCR LTER we ultimately chose to make minor modifications to our old system. We updated our version of EndNote so that we could use the expanded or improved features in the new version relative to linking to PDFs, generation of XML output and citation of electronic publications. We added a new management module to our Drupal installation that allowed us to rapidly, and completely, remove all the existing publications (one-by-one and even screen-by-screen removal of publications is unwieldy) so that a new “clean” import from EndNote could be performed. We use the unique Accession Number field to contain a “key” that can be used to link publications in the relational citation database created by Drupal with other data (e.g., personnel and datasets) that are in our relational database.

Good Reads

Review: Scientists Threatened by Demands to Share Data: the open data movement is polarizing the scientific community

edit

Review by Kristin Vanderbilt (SEV)

Scientists Threatened by Demands to Share Data: the open data movement is polarizing the scientific community
(http://america.aljazeera.com/articles/2013/10/10/scientists-threatenedbydemandstosharedata.html;
accessed on 5/21/2014)

The title of this short article promised drama but ultimately did not demonstrate that data sharing is causing a schism in the scientific community. Far from it.

In May 2012, the Obama Administration adopted open access policies that require data collected by federal agencies to be publicly accessible.   While some scientists applaud this change and believe that sharing information promotes scientific discovery, others fear that data sharing will lead to the discovery of errors in the data and theft of future research ideas.   This short article by science writer Victoria Schlesinger notes that the latter fears are sometimes realized.   Errors have been found and research ideas have been published without the input of the researcher who collected the original data.   A graduate student working in Peru, for instance, videotaped a new species of spider and shared his finding via Twitter almost immediately, thereby losing (he felt) an opportunity to publish in a top journal because someone beat him to it. To counter this threat of being “scooped”, ways to increase the rewards for data sharing are being sought.   The use of DOIs to make data citeable is noted, as are organizations that help scientists with the time-consuming chore of cleaning up and archiving data.

I found this article while looking for references on data sharing. Despite its title, the author presents little evidence that data sharing is “polarizing the scientific community.” Instead the author quotes several scientists who are supportive of open science in order to remain relevant and spur scientific discovery.   Even the grad student who was “scooped”.

Review: Troubleshooting Public Data Archiving: Suggestions to Increase Participation

edit

Mary Martin (HBR)

There is no shortage of diologue on the benefits and barriers to data archiving and data sharing. In this 'good read', of interest to researchers and information managers alike, Roche et al. (2014) propose a number strategies to increase particpation in public data archiving. They focus on four main areas where a number of challenges arise: data embargoes, communication between data providers and users, data-use ethics, and data recognition. Although they don't (and shouldn't) propose one-size-fits-all solutions to these issues, they make a number of points that are worth reading, and keeping in mind, as we continue to have these discussions within our research communities. In particular, the authors address the challenges that are unique to long-term datasets, and suggest strategies for publishers and funding agencies that may increase participation in data archiving.

Citation: Roche DG, Lanfear R, Binning SA, Haff TM, Schwanz LE, et al. (2014) Troubleshooting Public Data Archiving: Suggestions to Increase Participation. PLoS Biol 12(1): e1001779. doi:10.1371/journal.pbio.1001779