2013 March 11

Bridging Big Data Infrastructure

Summary of Workshop
4 – 6 December 2012
Taichung and Huisun Forest Station, Taiwan
http://event.nchc.org.tw/2012/datainfrastructure/index.php

Bridging Big Data Infrastructure

Summary of Workshop
4 – 6 December 2012
Taichung and Huisun Forest Station, Taiwan
http://event.nchc.org.tw/2012/datainfrastructure/index.php

From 4 – 6 December 2012 an interdisciplinary and international group of researchers met in Taiwan to discuss issues of Bridging Big Data Infrastructure. This report summarizes their discussions and potential follow-up activities.

This report has six sections:

A. Workshop Summary
B. Participant Statements to the Value of the Workshop
C. Potential Actionable Items
D. Short Report from Breakout Groups
E. Presentation Abstracts
F. List of Participants

There are three types of outcome for this workshop, which are described in the report:

1. Overarching Themes, common across the applications and the projects
2. Personal and Project Bridges, reflecting the interactions between participants
3. Actionable Items, outcomes identified across participants that could be achieved within an 18 month period, i.e., by June 2014

Because this workshop was organized by ongoing organizations (GLEON, PRAGMA, and NCHC) with interests directly in this area of bridging infrastructure, there is a strong motivation in ensuring many of the actionable items are realized.

We wish to gratefully acknowledge the support from the Taiwan National Science Council (NSC), the Taiwan National Applied Research Laboratory (NARL), and the US National Science Foundation (NSF). In addition, we thank the local organizer, NCHC, for providing the logistic support to allow participants to focus on the intellectual discussions and at the same time take advantage of the unique physical settings of the workshop. We also thank the Hui-sun Forest Station, of National Chung Hsing University, for allowing us the use of their facility in a very rich environment for discussions.

A. Workshop Summary

The science research architecture is changing dramatically in the beginning of the 21st century. Changes have resulted from investments in networking infrastructure; research in computing, data, and sensing; advances in technology and equipment developed and available for labs and observing systems; and deployments of observational networks and remote sensing systems. These investments have enabled a deluge of data (in volume, types, sources of data) by the scientific community, with the ensuing data intensive science of discovery often referred to as the fourth paradigm of science.

In addition, the conduct of science is changing, becoming increasingly collaborative, distributed, and international[1]. Many more groups have formed to work together across disciplinary, institutional, and geographic/political boundaries to solve larger community and societal challenges. This trend is leading to an increase of network science, the science conducted by a network of (self-organized) researchers, working with distributed data, models, resources, and people in the network.

Both the explosion of data available and the increase of these ad hoc and persistent networks of scientists are placing demands on the data infrastructure (i.e., data sets; metadata about sets; tools to contribute to, annotate, discover, share, track use of, and standardize data). Often the infrastructure is inadequate, poorly defined by the community of users, siloed (i.e. development in one community is not migrated to other communities), and thus inhibits scientific progress.

The Bridging Big Data Infrastructure (BBDI)[2] workshop brought together researchers from three distinct application communities that shared needs for addressing big or complex data[3] issues. Each community is developing infrastructure and does not often have the opportunity to communicate with the others; and each reflects the new conduct of distributed, network science. These three communities are:

- Limnology (process oriented, with emphasis on the ecosystem)
- Biodiversity (biologically oriented, with emphasis on organisms, distribution, and evolutionary processes)
- Disaster mitigation and recovery (event oriented, with emphasis on often real-time cross-system integration)

The workshop was organized by GLEON, the Global Lake Ecological Observatory Network; PRAGMA, the Pacific Rim Applications and Grid Middleware Assembly; and NCHC, the National Center for High-performance Computing of Taiwan. These organizations share a history of working together and making advances through partnering with other groups, focusing on pragmatic technology solutions to improve data integration across multiple data sources, and

[1] Royal Society. Knowledge, Networks and Nations: Global scientific collaboration in the 21st century. March 2011 (http://royalsociety.org/uploadedFiles/Royal_Society_Content/Influencing_Policy/Reports/2011-03-28-Knowledge-networks-nations.pdf)

[2] http://event.nchc.org.tw/2012/datainfrastructure/index.php.

[3] We use the NSF definition of "big data" as "large, diverse, complex, longitudinal, and/or distributed data sets generated from instruments, sensors, Internet transactions, email, video, click streams, and/or all other digital sources available today and in the future." See NSF 12-499.

making scientific and technical progress through defined expeditions that bring together multidisciplinary teams to address defined problems.

A major feature of the workshop was the participation of representatives of many projects or agencies, from the United States, Taiwan, Thailand, South Korea, and Germany, that share interests in addressing complex data issues in the environmental or geo-spatial settings. These projects[4] and groups include: DataNet: Data Observation Network for Earth (DataONE); National Ecological Observatory Network (NEON); DataNet: Sustainable Environment-Actionable Data (SEAD); Critical Zone Observatories (CZO); United States Geological Survey (USGS); US, Taiwan, and the East Asia Pacific Long-Term Ecological Research (LTER) Network; EarthCube; Research Data Alliance (RDAlliance); Taiwan Biodiversity Information Facility (TaiBIF); Taiwan Ocean Research Institute (TORI); and Taiwan's Disaster Management Information Platform. In addition, there was active participation from GLEON researchers, PRAGMA researchers and Expeditions on Biodiversity and Lake Eutrophication; and NCHC. Finally, there were participants from Thailand's NECTEC projects on disaster management and environmental data integration (E-RIUM: Environmental Informatorium).

There are two unique attributes to this workshop: the extent of diverse projects in a single meeting (room), and the desire to identify useful, tangible, and actionable approaches to collaborate, with vehicles (GLEON, PRAGMA, NCHC) to carry out some of the actions. In addition, the location of the workshop, at a neutral setting for all participants allowed for a great exchange of ideas. Furthermore, the international composition of workshop participants provided a global perspective on the three topical areas, and encouraged the participants to think of collaborating at a scale not always possible in domestic settings (groups collaborating internationally are not competing for the same pool of funds).

There are three products from this workshop. First, we discussed several overarching themes that we feel are relevant to community efforts to bridge big data infrastructure. Second, personal bridges were formed between projects and their participants. Finally, a set of potential, actionable items were identified that, if acted upon and accomplished, would set up longer term partnerships between participants, the projects these participants represent, and their communities.

Product 1: Overarching Themes

Among general common themes articulated at the workshop are the desire and opportunity to collaborate among groups and projects; the identification of shared technology challenges, including data discovery, data integration, assigning unique identifiers, provenance of data, and attribution; and the repeated statements that the culture of science and policy are as important as technology in influencing progress toward bridging data infrastructures. For example, both sharing of data and participation in network science (rather than of the individualistic approach to science of earlier decades) are as much cultural or policy issues as technological issues.

[4] We note that many of the US projects (e.g., DataONE, NEON, SEAD, CAO, LTER, RDAlliance, Earthcube, GLEON, PRAGMA) are supported in part by the US National Science Foundation and that many of those from Taiwan are supported by the National Science Council or the National Applied Research Laboratory. This may provide opportunities for work between projects.

Three specific common themes that emerged from the discussions include:

- Workflows (distinct form workflow software) are increasingly important in the conduct of science in a data-rich world. The practice of harnessing complex data will be enhanced by understanding workflows, and then developing and using infrastructure to automate the processing of information and the conduct of research.
- Training in areas of data and use of data infrastructure (including software and workflows) is an immediate need. Knowing what skill sets is available is a critical first step. As new analytic tools become available, new applications will leverage that training.
- Scientific questions, the underlying infrastructure (including software) and technology development, and training are three interwoven components of science practice. Science must drive some technology development and infrastructure deployment; training must be addressed in the context of the science and the available (and future) infrastructure; infrastructure (and data produced) will allow new types of questions to be asked.

Product 2: Personal and Project Bridges

Many of the participants stated (See Appendix on Perspectives on Community Interaction by individual participants) that a value to the workshop was the personal connections made with individuals with common interests. Two examples of potential activities between projects and groups are given here. Others are listed in the report.

- Creating a DataONE node by GLEON. The value of this activity is to provide a vehicle to assign unique identifiers to GLEON data, to have the data published, and to track use of the data. In addition, DataONE will learn from the ease or difficulty GLEON experiences in becoming a member node. This activity also has the potential to engage PRAGMA technologies of packaging, developed by NCHC and UC San Diego.
- Training and workflow. One action item is to review GLEON working group activities over the last several years, and identify challenges of discovery, integration, and analysis of data. This will lead to both identifying several possible workflows and identify skills that could be developed to better accomplish data analysis across multiple lakes. This activity has the potential of working with NEON in identifying skills for network science. In addition, this activity has the potential for engaging the GLEON Student Association and the PRAGMA Students group.
- Unique identifiers. Many groups in the biodiversity and environmental communities are discussing how to uniquely tag or "identify" data. One follow-up is to bring together groups of the meeting with activities of BiSciCol (Biological Science Collections, http://biscicol.blogspot.com/) and iDigBio (Integrated Digitize Biocollections, https://www.idigbio.org/); another is to look at what other groups are doing, such as RDAlliance, TDWG (Taxanomic Database Working Group, www.tdwg.org/), CDL (California Digital Library, www.cdlib.org/), USGS (United States Geological Survey). [One outcome would be to recommend to GLEON of which identifiers to develop linked data.]

Product 3: Actionable Items

Over the next 12 to 18 months there are several opportunities for subsets of the participants to guide and observe progress from this workshop. GLEON 15 will take place on 4 – 8 November 2013, in Bahia Blanca, Argentina. PRAGMA 24 will take place 19 – 22 March 2013 in Bangkok, Thailand; PRAGMA 25 will take place in mid-October 2013 in Beijing, China, and PRAGMA

26 is scheduled for Spring 2014. NCHC is planning to host the Southeast Asia Institute Program in December 2013, and the Research Data Alliance plans to have recommendations for universal identifiers by Summer 2014. As noted above and in the report, several of the concrete actions can take place at these or other meetings. For example, DataONE could conduct training at a GLEON workshop, or PRAGMA could package tools for DataONE and GLEON to be used by GLEON members. Similarly there will be follow-up on unique identifiers through initially convening groups of DataONE, RDAlliance, and BiSciCol, either in a PRAGMA workshop or at a PRAGMA related event.

B. Participant Statements

In this section many participants responded to the question of what they gained by participating in this workshop.

Reed Beaman, University of Florida
I was able to connect to other experts and learn more about efforts in networking biodiversity information, including biological collections and taxonomic data, in Taiwan and Southeast Asia through the efforts of TAIBIF. We also discussed challenges in extending efforts and success to developing countries in the region. It was also an opportunity to better understand how the HPC solutions can and are applied to domains with which I am less familiar. The challenges for data integration and access are familiar ones, but it was somehow reassuring to know that other domains face similar issues in archiving, interoperability, provenance, metadata, and data publication. Many of the emerging solutions discussed, e.g., within the DataNets and Research Data Alliance, parallel those in the biodiversity community (a DataNet stakeholder). It was also apparent that we can do much more as overlapping international communities and domains to share solutions, so bridging big data efforts need to continue an emphasis on collaboration across the socio-political gaps. In a breakout session on data integration and discovery, we agreed that basic recommendations on persistent, resolvable, global unique identifiers (GUIDs) could be achieved as a short-term outcome.

Cayelan Carey, University of Wisconsin
As a limnologist, this workshop was extremely valuable in exposing me to new technologies for data access, management, and curation- this information is currently not readily available for ecologists 'on the ground'; i.e., analyzing large, complex, and messy datasets, and I learned about several new tools that I plan to use in my research (e.g., Kepler, Morpho, and DataOne's new Excel program, DataUp). Perhaps more importantly, I made several great connections in the data infrastructure community, and hope to use this network to facilitate new data-sharing opportunities for GLEON within the next few years.

Chris Duffy, The Pennsylvania State University
The workshop was extremely successful in exploring new ways to deal with the problem of "BigData" in contemporary ecological sciences. The workshop defined BigData in a way that goes beyond data that is collected by individual scientists, to include the larger data that serves to

support the sharing of diverse data resources within and across disciplines. There were important discussions of BigData that provides the context for scaling up individual hypothesis-driven science, small-group research as well as team-science observatory research, to a global context. There were useful discussions about improving the community conceptions of typical "use cases" and developing a more formal strategy to deal with the various workflows that GLEON/NEON/CZO/etc scientists use, or might use to advance international science, and global science research. It was very impressive to me the level of formal/informal science organization and community science development that GLEON has achieved without core funding. It was also interesting that GLEON was serving the role of the international/global extension of NEON to advance global aquatic ecosystem science. Similar discussions are happening within the Critical Zone community funded by NSF and international partners interested in critical zone science globally.

Perhaps one of the most important things I took away form the meeting was the critical role that recent PhD Post Doctoral scientists were playing in making GLEON happen. This investment in the work force of science through development of "expeditions" with Post Doctoral leadership is an important strength of the effort.

Corinna Gries, University of Wisconsin
Planning concrete activities involving several projects, leveraging developments, talents and resources was the most valuable for me.

Paul Hanson, University of Wisconsin
One general observation is that CI scientists and domain scientists are working at a level of cooperation that's well beyond our previous experience. In the past, collaborations of this sort often began with domain scientists talking about what they want and CI scientists talking about what they had already developed and might apply. In the past, the problem has been trying to fit the science into the CI. At this meeting there seemed to be a qualitatively new level of cooperation and synergy – one in which both CI and domain scientists really are working at addressing solutions relevant to the domain scientists. This does not necessarily mean solutions come easily, just that the processes is not unnecessarily side-tracked by solutions looking for a problem.
There also were serendipitous outcomes of this workshop, including:

- Duffy (CZO)-Hanson (CDI) workshop
- Read (USGS)-Duffy (CZO) better understanding:
- Improvement in DIBBs planning and activities
- DataONE test runs

Tim Kratz, University of Wisconsin
This was the first meeting where people representing GLEON, PRAGMA, NEON, DataONE, LTER, CZO, USGS, NCHC, TFRI and perhaps others were able to talk about common problems and, more important, solutions to data issues. One tangible outcome will be to link GLEON with DataONE by making GLEON a tier 3 DataONE member node. We also learned how best we might interact with NEON, CZOs etc. It was a very pragmatic meeting and highly useful.

Chau Chin Lin, Taiwan Forest Research Institute
It was a great experience to participate this workshop. The workshop titled "big data" with many related project presented from lake, biodiversity and disaster domains of people. All the presentation showed that big data means not only large volume of data but also diverse and complex of data. In this background view point, I learned that collaboration under a CI platform and kinds of standards are necessary for integrating big data in terms of volume and heterogeneous do data sources. In addition, creating of standard workflows to serve for data analysis is also necessary in the future.

Jordan Read, United States Geological Survey
Personally, I had several conversations that were extremely beneficial to potential future work/collaborations. USGS CIDA has played the role of primary software developer for the USGS's BioData (aquatic.biodata.usgs.gov), which is a database and retrieval system for complex biological assessment data. This system is potentially applicable as a solution to some of LTER and NEON's complex data storage and retrieval needs. Regardless of future outcomes, it was a good use of time to share some of the details of this project and overlap with other initiatives, and to establish a dialogue with other data scientists.

Sornthep Vannarat, National Electronics and Computer Technology Center (NECTEC)
Participating in this workshop was a great opportunity. It was very fascinating to learn about Big Data projects e.g. DataONE, NEON, GLEON, and, CZO; and, how these projects work together to tackle Big Data challenges. Chris Duffy's presentation of Cyberinfrastructure for Watershed-scale Prediction gave me an important idea that data can be used as a platform to stimulate and direct research. I also have a clearer thought about data integration. Of course, there are still a lot of issues that must be resolved. At least, we have discussed some of them in this workshop such as data ownership and benefit of data sharing to the data owner; security, integrity and privacy; global identifier and data discovery. This workshop gave me a chance to learn from many experienced investigators. I learned about some standards and tools for data integration, such as WaterML and MetaCat, that will be useful for my current projects.

Brian Wee, National Ecological Observatory Network
As part of NEON's engagement with other NSF funded science networks, we have undertaken a number of exploratory activities to determine areas of alignment with LTER. These activities included a NEON-LTER co-authored article on the NEON website (http://goo.gl/IKT7l), a NEON-LTER session at the 2012 LTER All-Scientists' Meeting, and a panel discussion comprising LTER, OBFS, and USA NPN leaders at the 2012 NEON Inc Annual Membership meeting. As a result of the PRAGMA/GLEON workshop, I see opportunities to pursue similar joint activities with GLEON. One immediate area of potential activities is the involvement of the GLEON community in the review of the NEON protocols that are currently under development. These will be released for public review and comment in 2013. At the LTER/OBFS/NPN/NEON panel discussion in October 2012, participants discussed the latent demand for standardized NEON terrestrial biological at LTER sites. Successful deployment of

such protocols would enable enhanced harmonization between LETER and NEON field measurements. This would lend more power to multi-site generalizations. Similar ideas should be floated before the GLEON community.

Another outcome attributed to this PRAGMA/GLEON workshop was the discussion to identify the training needs of our respective (GLEON, NEON) communities in the area of data management and informatics. I have been involved in other discussions with the ecological informatics community about similar issues, and there has been an explicit demand for such training from NEON, Inc.'s dues-paying member institutions. This workshop advanced those discussions by suggesting that we frame those informatics training needs within the framework of a scientific workflow. After such workflows are documented, one can determine how emerging data management and informatics practices can be used to facilitate, improve, and accelerate such workflow processes. There also already various training needs and training packages identified from other workshops, and the documented workflows can be used to verify the completeness of such training needs.

C. Potential Actionable Items

In this section most of the actionable items discussed at the workshop are listed. The idea is that these items could be accomplished in the next 18 months, that is by June 2014. Many are proposed and detailed in the reports from the breakout groups. Others were proposed in plenary session or in other small groups.

Establishing and Gaining Facility with DataONE Technology

- GLEON will become a DataONE tier 3 member node and provide access to at least one GLEON value-added, derived dataset.
- In addition GLEON will convene two focus groups to assess the ease of use of the DataONE system by GLEON scientists.
- See additional actions and expanded descriptions in DataONE section below.

Scientific Workflow and Training

- Document workflows used for comparative lake analysis in GLEON. Use results on workflows to
    - Develop training skills
    - Automate some processes
    - Note: The term workflow is use to describe the scientific process of getting a result. In this case it is used to understand all of the steps required to create a publication from integrating data from various data set.
    - Note: See report of breakout group on Scientific Workflows and Training for details.
- Have two groups in GLEON try to execute comparative lake workflow, one using "traditional" approach, other trying to use technology
    - Both cases feedback to improving technology

- Invite CI group(s) to GLEON workshop to provide tutorial or overview

Data Integration: Short term steps (6-12 month) for unique identifiers

- Review of existing identifiers uses and the key stakeholders (DOI organization, RDA, TDWG, CDL, USGS)
- Develop recommendations for identifier use for different groups, e.g., DOIs for published materials, etc.
- Develop recommendations for how to use the identifiers to develop linked data. For example - a GLEON site should have a number of base layers that should be tied to the identifier for the location / site. Samples etc taken at the location should also reference the site identifier as well as include their own unique identifier per sample.
- BiSciCol hosting a meeting Jan 7-8, 2013 with iDigBio. Major topic is application of identifies to BioCollections. Invite for an open tele/video conference discussion session (ca. 2 hours) can be extended to participants of the Taiwan workshop. (Note: Rather than the original though to hold this meeting in early January, the group will reschedule this so part of the follow-up discussion can happen at a PRAGMA workshop or related event.)

Other Potential Activities

- Packaging DataONE Software
    - Take advantage of PRAGMA experience in virtual machine images and creating reproducible software
- Trust envelop experiment (i.e., creating a virtual private network for collaboration amount distributed sites, including data and compute resources and the team members)
    - Will be tried in PRAGMA Biodiversity Expedition for sharing sensitive data
    - Can be used to better execute transporting virtual machine images for multiple parameter jobs
- Provenance experiment (tracking who is using data)
    - Will be tried in PRAGMA Biodiversity Expedition
    - Can be used to track which parameters uses in simulation experiments

DataONE Section

Four actions for followup that emerged directly from the Taiwan workshop:

1. There are several software products being developed on the DataONE project that may be generally applicable to a wide audience of researchers, particularly those engaged in the environmental sciences. To ensure that such software products are readily deployable it is critical that they are packaged so that installation is as straight forward as possible. The software packaging expertise available through the PRAGMA group is an opportunity that could benefit both PRAGMA through diversifying the suite of software packages available, and DataONE by streamlining the software distribution process. Beginning early 2013, DataONE will work with PRAGMA to develop a distribution package for a DataONE Member Node software stack

(selecting either Metacat or the "Generic Member Node", a reference implementation of a Member Node software stack) to simplify installation as far as practicable. In addition, DataONE and PRAGMA will investigate funding opportunities to support the packaging of additional components, in particular elements of the DataONE "Investigator Toolkit" with the general goal of developing a suite of easily installable, cross platform tools that address all aspects of the data lifecycle. Dave Vieglais of DataONE and Phil Papadopoulos of PRAGMA will lead the activities.

2. Corinna Gries has a pending proposal to the NSF DIBBS solicitation which, among other activities, plans to utilize DataONE infrastructure to facilitate long term preservation and access to content generated from the GLEON partners activity. As a first step towards these goals, DataONE will work with GLEON to install at least one Member Node (likely based on Metacat) to start evaluating the benefits of participation in the DataONE federation and to help streamline the process in anticipation of a wider installation base across multiple GLEON locations. This activity provides a real-world use case requiring support from action #1 above, and will provide usability feedback on products packaged from that activity and others. Dave Vieglais of DataONE, Corinna Gries and Cayelan Carey of GLEON will lead the activities.

3. Interest in participation with DataONE by way of Member Node installation was expressed by a number of participants at the workshop. DataONE will continue to work with various representatives such as Chau-Chin Lin for the Taiwan Forestry Research Institute to ensure that Member Node deployment proceeds in an orderly fashion. This activity leverages outcomes from both #1 and #2 above, and adds an additional requirement of multi-lingual support for any installation procedures and documentation.

4. Discovery of content across diverse holdings in DataONE is achieved by generation of a common search index populated by metadata describing datasets that are retrieved from Member Nodes participating in the DataONE federation. A key challenge as the breadth of participation expands beyond locations with English as the primary language is how to ensure that relevant content remains discoverable to all researchers regardless of their language of choice. One approach that addresses this goal is to develop multi-lingual mappings to the conceptual equivalent of controlled vocabularies and taxonomies, and linking indexed documents to the concept identifier rather than the literal value. In doing so, user interfaces and query expressions may be developed in the native language of the researcher, translated to the appropriate mapping by the query engine, and matching documents selected based on presences of the concept identifiers. Implementing such a process requires selection of appropriate vocabularies and taxonomies, generation of identifiers for all the concepts expressed therein, and generating translations to the various languages to be supported. The indexing process will also need to be updated and appropriate multi-lingual user interfaces developed to enable access. The breadth of activity for this project indicates additional support will be required, and as such the first action will be to develop a proposal to acquire funding to support the activity.

D. Short Reports from Breakout Groups

One of the goals of the workshop was to find commonality of approaches across three driving application areas. Based on plenary discussions of the entire group, several topics were identified

that interested at mix of application and technology projects and areas, and for which there were tractable and actionable items over the course of a subsequent 18 month period. Three groups identified items:

- DataONE and GLEON
- Scientific Workflows and Training
- Data Integration

A summary of their discussions and actionable items are contained in this appendix. A fourth group, discussing the topic "What to scientists want?" did not have sufficient time to conclude discussions.

GLEON and DataONE

Participants: Tim Kratz (U Wisconsin, Chair Breakout Group), Dave Vieglais (DataONE), Corinna Gries (U Wisconsin), Cayelan Cary (U Wisconsin), Philip Papadopoulos (UCSD), Beth Plale (Indiana U), Hsiu-Mei Chou (NCHC), Brian Wee (NEON)

Issue: GLEON has several kinds of datasets, including streaming sensor data and static, value-added datasets. GLEON is currently facing the challenge of how to make static value-added datasets discoverable and accessible while heeding original data providers conditions and providing proper attribution and use tracking.

Actionable items: GLEON will become a DataONE tier 3 member node and provide access to at least one GLEON value-added, derived dataset. In addition GLEON will convene two focus groups to assess the ease of use of the DataONE system by GLEON scientists. We believe this can be accomplished by the "GLEON DIBBS team" within 6-12 months.

Benefits: Datasets in DataONE are given digital object identifiers and are citable, giving appropriate attribution to the dataset contributors (both the original data providers and those whose work added value to the derived dataset). DataONE also provides a mechanism to control access to selected subsets of data to meet the original conditions of the data providers. Finally DataONE tracks downloads of datasets allowing the data providers information on data reuse.

Scientific Workflows and Training

Participants: Jordan Read, (USGS, Chair Breakout Group), Cayelan Carey (U Wisconsin), Brian Wee (NEON), Tim Kratz (U Wisconsin), Paul Hanson (U Wisconsin), Apivadee Piyatumrong (NECTEC), Hsiu-Mei Chou (NCHC), Hen-Biau King (formerly TFRI and ILTER), Peter Arzberger (UCSD)

This breakout session discussed the category of scientific workflows, and how these workflows may be used to identify clear areas where additional scientific, analytic, technological, social, and communication training may be necessary. The group was specifically interested in the scientific approach of collaborative groups attempting to answer science questions that are only made possible by leveraging the expertise and diverse data (multi-site)

that are found in science networks. Within the broad category of network science, the group focused on grassroots networks, which have often been shown to be effective and efficient scientific contributors.

In an international science landscape of increasing data volume and complexity, we are interested in identifying specific limitations to the productivity of grassroots science networks. These "bottlenecks" to network science may be grouped within the broad categories of communication, expertise, data quality/availability, and software/hardware needs. By identifying common bottlenecks, we can encourage these groups to engage external participants (if applicable), or suggest specific scientific training. In order to provide a basic assessment of the potential needs of grassroots network science groups, we plan to carry out the following tasks:

1. Establish a list of known historical multi-site science projects within GLEON (GLEON providing the cases studies because documentation exists for many of these projects, and the network provides a good cross-section of international collaborative science). These projects should represent a wide distribution of based on a definition of success criteria, ranging from projects that achieved all desired goals, to those without momentum to continue beyond initial phases.
2. The group will meet once for multiple days to establish a survey that will be used for assessing the projects identified in step 1. The group will send out the survey to the project participants at the end of this meeting. Efforts will be made to include social scientists in this and future workshops.
3. The group will meet a second time, with the purpose of analyzing survey results and drafting a publication.
4. The group will finalize and submit the publication, and potentially participate in other national/international meetings to share the results via oral presentations.

Data Integration

Participants: Reed Beaman (U Florida, ,Chair Breakout Group), Dave Vieglais, (U, Kansas, and DataOne), Philip Papadopoulos (UCSD), Sornthep Vannarat (NECTEC), Bo Chen (NSPO), Whey-Fone Tsai (NCHC), Shyi-Ching Lin (NCHC), Fang-Pang Lin (NCHC), Hsiu-Mei Chou (NCHC), Beth Plale (Indiana U, and Research Data Alliance), Chris Duffy (Penn State U, and Critical Zone Observatories), Corinna Gries (U Wisconsin, and Long Term Ecological Research), Kwang-Tsao Shao, Academia Sinica, and Taiwan Biodiversity Information Facility)

Overview of Breakout Group Discussions: The overarching topic of this breakout group was data integration, with secondary topics on data discovery and persistent global unique identifiers.

Integration: One of the key issues of current data intensive science is the ability to integrate data from multiple sites, and of multiple types. Although a broader issue than can be tackled by members of this workshop, examples of data integration issues in the context of the represented science domains were discussed that would benefit from infrastructure and technologies that facilitate integration. The examples represented heterogeneous and complex data sets stored in multiple repositories, which can empower research and downstream applications when through strategically integrated and/or bundled. Geospatial, in particular earth observation data, was emphasized for its broad applicability. Domain specific issues are outlined below:

- GLEON: There is a need to integrate climate, geographic/geospatial data, biodiversity information, along with buoy and satellite data, to be incorporated into models. The globally distributed buoys ultimately represent a sensor network. Initial efforts will be in establishing scalable computational capability to support modeling efforts that can in turn be applied to a growing database.
- Biodiversity: Rapid growth of digital data generated in biodiversity science is characterized by a heterogeneous array of taxonomic, genetic, functional and ecological information collected in the field and in the lab, as well as a need to publish data products resulting from analyses. While a large component reflects improved DNA sequencing technology and lower costs, there are other data drivers, such as digital imaging products. There is also a further global need to digitize and integrate data from biological collections, which represent physical and permanent documentation of life on Earth. Several efforts were represented at this workshop (TaiBIF, iDigBio, DataONE), and several efforts were cited, e.g. Global Names Architecture (GNA), Genomic Standards Consortium (GSC), Biocode Commons, GenBank, LTER and NEON.
- Disaster management: Disaster management infrastructure carries with it the need to respond to crises quickly with appropriate data. The use of bundled data products in this community were viewed as exemplary. Large data sets, e.g., collation of environmental layers (C. Duffy, USGS), could be accessed effectively and efficiently for specific responses (e.g., earthquakes) and geographic areas.

In addition, several approaches were discussed that are potentially relevant to many areas of science to address aspects of integration

- Metadata capture (EML, Metacat, Dublin Core). DataNets (DataONE, SEAD) are beginning to address basic integration issues, especially at the dataset level.
- National and regional approaches bring some spatial bound data together, e.g.,Taiwan Interagency committee on biodiversity data integration.
- Trust envelopes (overlay networks), especially for collaboration prior to public data publishing and/or to address sensitive, proprietary, or regionally specific data sets.
- Approaches based on geospatial data, e.g Open Geospatial Consortium standards, GEOBON and regional BONs relevant (e.g., AP BON), scalable projections (e.g., Military Grid Reference System MGRS).

Discovery: Data discovery is often considered the first step in data integration – asking what is available, in what format, and what is appropriate to integrate. The group looked most closely at the effort by C. Duffy of USGS described above, where aggregated content relevant to some purpose is bundled to improve access. One core challenge that was cited is how certain queries handle / support multiple languages, and what that would mean to capturing metadata.

Identifiers: The core issue discussed by the group was the question, "are persistent, resolvable globally unique identifiers (e.g., DOIs, ARKs) a prerequisite to a functional, integrated big data infrastructure where storage, archiving, and access is scalable and reliable?" This issue is being discussed by several groups now, such as GBIF, DataONE, Research Data Alliance (RDA), IDigBio,BiSciCol, and the geospatial community, e.g. Open Geospatial Consortium. The RDA plans to study this issue and have a report in roughly mid to late 2014.

Short term steps (6-12 month):

1. Identifier stakeholder organizations including Research Data Alliance (RDA), Taxonomic Data Working Group (TDWG), California Digital Library (CDL) United States Geological Survey (USGS), iDigBio, GBIF.
2. Develop recommendations for identifier use for different applications, e.g., DOIs, ARKs for published materials.
3. Develop recommendations for how to use the identifiers to support/enable linked data applications. For example - a GLEON site should have a number of base layers that should be tied to the identifier for the location / site. Samples etc taken at the location should also reference the site identifier as well as include their own unique identifier per sample.
4. Hold a meeting of BiSciCol, iDigBio, and other participates at a PRAGMA workshop or other relevant event. An early topic is application of identifies to biocollections.

Given the amount of time at this workshop, the breakout group limited itself to identifying the above list of short-term actionable items relevant to identifiers. One of the priorities should be mutual, broader awareness of stakeholders, as needs and interest are cross-cutting. There was also some enthusiasm about bundling and creating a data resource for global lakes. This is worth future consideration.

E. Presentation Abstracts

List of Abstracts (by Presenter)

Reed **Beaman**, University of Florida, *Documenting Biodiversity: Infrastructure-enabled Research in Unusual Environments*

Cayelan **Carey**, University of Wisconsin, *GLEON-enabled Science: A Research Sampler and Case Study of the Data-driven Approach to Global Limnology*

Chris **Duffy**, The Pennsylvania State University, *Towards a Cyber infrastructure in Support of Big-Data for National and Global Ecohydrologic Sciences*

Corinna **Gries**, University of Wisconsin, *LTER Information Management Overview* and *Next steps for Data Sharing in GLEON*

Paul **Hanson**, University of Wisconsin Center for Limnology, *GLEON Cyber-infrastructure Successes and Unmet Challenges*

Tim **Kratz**, University of Wisconsin, *GLEON Background Presentation*

Chau-Chin **Lin**, Taiwan Forest Research Institute, *Development of Ecoinformatics in EAP-ILTER*

Philip **Papadopoulos** and Peter **Arzberger**, University of California San Diego, *PRAGMA Overview and Directions*

Apivadee **Piyatumrong**, National Electronics and Computer Technology Center (NECTEC), *E-Rium: Environmental Informatorium*

Jordan **Read**, United States Geological Survey (USGS), *USGS Activities*

Kwang-Tsao **Shau,** KC Lai, YC Lin, CH Hsu, HY Li, LS Chen, GS Mai, H Lee, Biodiversity Research Center, Academia Sinica, *International Cooperation is the Key to the Success of Taiwan's Biodiversity Information Integration*

Whey-Fone **Tsai**, National Center for High-performance Computing (NCHC), Bo Chen (NSPO), Chi **Wu** (TORI), Sornthep **Vannarat** (NECTEC), Shih-Ching **Lin** (NCHC), *Case Studies and Needs: Disaster Mitigation*

Sornthep **Vannarat**, National Electronics and Computer Technology Center (NECTEC), *NECTEC and Disaster Mitigation Projects*

Dave **Vieglais**, University of Kansas and DataONE, *About DataONE*

Brian **Wee**, NEON, *Overview of the National Ecological Observatory Network*

Introduction to Abstracts

This set of abstracts reflects most of the talks presented at the workshop. There are several categories of talks:

- Case studies that describe needs in the realm of big data. The key application areas include lake ecology (Hanson, Carey), Biodiversity (Beaman, Shao *et al*), Disaster Management (Tsai, Vannarat).
- Existing networks that either collect, manage data or develop tools to analyze data, such as GLEON (Kratz), LTER (Gries), the East Asia Pacific ILTER (Lin), Critical Zone Observatories (Duffy), USGS (Read), and NEON (Wee); or help develop cyberinfrastructure tools by working closely with application groups, such as PRAGMA (Papadopoulos)
- Projects that develop technology, such as E-Rium (Piyatumrong) or DataONE (Vieglais)

The abstracts below are arranged by lead authors family name.

Abstracts

Reed **Beaman**, University of Florida, *Documenting Biodiversity: Infrastructure-enabled Research in Unusual Environments*

Biodiversity science, broadly defined, intersects the study of molecular biology, genetics, organismic and taxonomic relationships, and ecology, addressing the need to understand complex interactions at multiple scales. Data, tools, and infrastructure that add to the biodiversity knowledge base are increasingly important to basic and applied research, such as management of natural resources. The loss of biodiversity is a global environmental issue, and addressing the technical challenges requires multi-disciplinary, international effort. This presentation described a project-level case study about a global biodiversity hotspot: Mount Kinabalu (4095 m.), Sabah, Malaysia, which is known for both animal and plant diversity (over 5000 vascular plant species), precipitous topography, and unusual ecosystems, including ultramafic habitats. An interoperability experiment was designed in a PRAGMA-facilitated collaboration that examines the scientific, functional, and technical requirements for integrating and analyzing data from exisiting biological collections, especially data on organism distribution, remote sensing, and climatic data. Further plans for deploying HPC resources, including a virtual cluster within a trust overlay network are in progress. The unusual and extreme habitats of Kinabalu provide unique opportunities to examine both the biodiversity and address how scalable infrastructure can address science challenges.

Cayelan **Carey**, University of Wisconsin, *GLEON-enabled science: A Research Sampler and Case Study of the Data-driven Approach to Global Limnology*

During the past seven years, there have been substantial research accomplishments resulting from GLEON's grassroots scientific working groups. The early GLEON projects typically

focused on one lake and were data-driven; i.e., they leveraged GLEON's sensor networks to document observed patterns in lake physics or chemistry. More recently, GLEON working groups have advanced to cross-site, question-driven analyses involving comparative studies of many lakes. In addition, these newer projects have also included biological datasets, analyzing trends and interactions between microbes and plankton. Examples include analyses of typhoon forcing of lake mixing and microbes (Shade et al. 2010a,b), long-term precipitation effects on water clarity (Gaiser et al. 2009a,b), the development of an automated metric for determining changes to ice cover in northern hemisphere lakes (Pierson et al. 2011), hurricane effects on thermal stratification in nine lakes (Klug et al. 2012), observing high-resolution variability in dissolved oxygen concentrations across 25 different systems (Solomon et al., Accepted), determining the effects of high-frequency changes on phytoplankton diversity in 5 lakes (Muraoka et al, In preparation), as well as many others. As GLEON makes the transition from data-driven to question-driven science, which has primarily been fueled by new tools and technologies that have facilitated large dataset analyses, a number of challenges and opportunities have arisen. For example, many working groups have used traditional tools for their research, such as emailed Excel spreadsheets and individual (non-scripted) QA/QC approaches; many working groups have struggled with authorship issues because not all datasets are created equal, and there have been difficulties in finding and accessing data needed for a project. As a result, working groups have identified several best practices, including: reiterating attribution policies with data providers throughout the analysis process; adopting universal data standards (e.g., controlled vocabulary, standardized workflows, QA/QC processes) throughout the network, new ways to archive and share curated datasets, and the need for more training to ensure that all group members are able to find, collect, manage, and analyze large datasets. As these best practices are developed and implemented, GLEON working groups will be able to further transition from question-driven to predictive science.

Christopher **Duffy**, The Pennsylvania State University, *Towards a Cyber infrastructure in Support of Big-Data for National and Global Ecohydrologic Sciences*

Abstract: This paper discusses a prototype cyber infrastructure that provides researchers, educators and resource managers with seamless access to essential geospatial/geotemporal data, computational models, and the data fusion tools necessary to support and advance ecosystem science. The evaluation of ecosystem and watershed services such as the detection and attribution of the impact of climatic or landuse change are examples of the pressing need for high resolution, spatially explicit resource assessments. However, the current cyber infrastructure for supporting models and data at a continental and global scale must overcome important problems of: 1) The accessibility of high resolution (space and time) terrestrial data sets. 2) Harmonization of essential environmental variables from multiple government agencies, countries and disciplinary science institutions. 3) Scalability of BigData in support of diverse scientific use cases that range from “long tail” scientists to high performance computation users for modeling, analytics and visualization of BigData. In this context I discuss a prototype hardware and software for web access to essential terrestrial data processing in support of catchment hydrologic and isotope modeling, and propose a strategy for consuming this data within a framework that enables ecohydrological modelers to build and test models using cyber

infrastructure with fast data access. Given the prospect of petabytes of existing environmental data, our prototype begins with a set of "essential terrestrial variables" (ETV's) necessary to provide the first level of support for model implementation anywhere in the continental USA. This challenge focuses on creating tools for fast data storage and access, as well as the computational resources necessary for high resolution, watershed simulations using models such as the Penn State Integrated Hydrologic Model (PIHM), Biome-BGC, and others.

Corinna **Gries,** University of Wisconsin, *LTER Information Management Overview*

The US Long-Term Ecological Research (LTER) network is funded by the National Science Foundation and was initiated over 30 years ago. The 26 LTER sites have had the mandate to archive their data in a meaningful way since their inception. Most sites generate and use several different types of data. Long-term core data are usually collected with the same methods at certain time intervals. Data from specific projects or experiments are generally shorter in duration and pose specific information management challenges due to their complexity in manipulation and sampling approaches. Streaming sensor data have been added more recently and although very uniform, are generating large volumes of data. Spatial data are widely used and generated by LTER sites. During these past 30 years LTER Information Management (IM) approaches have changed and adapted to emerging technologies, with the current expectation that all LTER datasets are well documented in the Ecological Metadata Language (EML) and published on-line, accessible to the general public. The LTER network currently has well over 6000 metadata documents publicly accessible via its data portal (https://metacat.lternet.edu). Due to the long history of Information Management, approaches have developed according to site specific needs and each site supports the full the data life cycle (collection, quality control, metadata generation, archiving/preserving, public access, data integration and data analysis) in slightly different ways. The new LTER Network Information System (NIS) will centralize access to LTER site data by harvesting site provided EML documents, retrieving the data as described by the EML and ingesting both into the Provenance Aware Synthesis Tracking Architecture (PASTA) framework. Data will be accessible via a LTER data portal and programmatically through webserivces. LTER is a member node of the DataONE (https://www.dataone.org/) and all data will be available through their data search as well. The new LTER Network Information System will be launched in early 2013.

Corinna **Gries**, University of Wisconsin, *Next steps for Data Sharing in GLEON*

The Global Lakes Ecological Observatory (GLEON) is a grassroots network of limnologists, ecologists, information technology experts, and engineers who have a common goal of building a scalable, persistent network of lake ecological observatories. This international community came together under the premise that data from these observatories will provide a better understanding of key processes such as the effects of climate and landuse change on lake function, the role of episodic events such as typhoons in resetting lake dynamics, and carbon cycling within lakes. Therefore, the basic idea of sharing data is well accepted while many cultural differences, funder requirements, career advancement considerations, and IT support/skill levels need to be taken

into account. Regular GLEON meetings have advanced the people network and expanded collaborations. Data sharing is currently accomplished through this people network, based on who knows who and insider knowledge of existing data sets. Early attempts in employing technology to automate the process of data sharing and especially making data better discoverable were only moderately successful and an IT taskforce was charged with making recommendations for approaches addressing the various needs and obstacles in data sharing. Two data 'types' have been identified each with its own set of requirements and challenges. Streaming sensor data are in high volume and need to be quality controlled in near real time, stored and made accessible in a user-friendly format. That is, most users don't need all variables for all times and sites and prefer to work with a specific, user defined subset of this large data volume. Value added data products are produced by an individual or a working group by aggregating and harmonizing data from many sources. This is still a labor intensive undertaking and these datasets are extremely valuable well beyond their original research purpose. Therefore, a system is needed that ensures appropriate credit to the original author(s) when the data are re-used in another project.

It is now proposed to collaborate with DataTurbine, DataONE and CUAHSI to address some of the outlined requirements for a more automated data sharing system. By becoming a DataONE member node and making the full software stack developed by this project available to GLEON participants, value added data products may be archived with appropriate access control. In addition DataONE will assign a Digital Object Identifier (DOI) which may be used for citing the dataset in subsequent publications, therefore crediting the original author(s) appropriately. It will be possible to discover data through the DataONE interface. The GLEON community will be introduced to the technology and extensive user testing will provide valuable feedback to the DataONE project. Similarly, the CUAHSI software stack in combination with DataTurbine will be introduced to the GLEON community and tested for managing streaming sensor data. Both approaches will not involve writing of new software, but existing software will be installed centrally and GLEON members will be trained in their use and will evaluate them with regards to technical skill levels required, usability, and sustainability within GLEON, i.e. will people be able and willing to use them when back at their home institution.

Paul **Hanson**, University of Wisconsin Center for Limnology, *GLEON cyber-infrastructure successes and unmet challenges*

National and global ecological issues transcend traditional boundaries, and scientists are eager to expand beyond traditional boundaries of geography, disciplines. This is driving ecologists to seek collaborations within the field that result in bigger and more complex data sets. The Global Lake Ecological Observatory Network (GLEON) has nearly 10 years of international collaboration on lake ecology. A long-term goal of the organization is to expand traditional science from the local scale centered in the discipline of aquatic ecology to a more global scale that crosses ecosystem boarders and disciplines (Fig. 1).

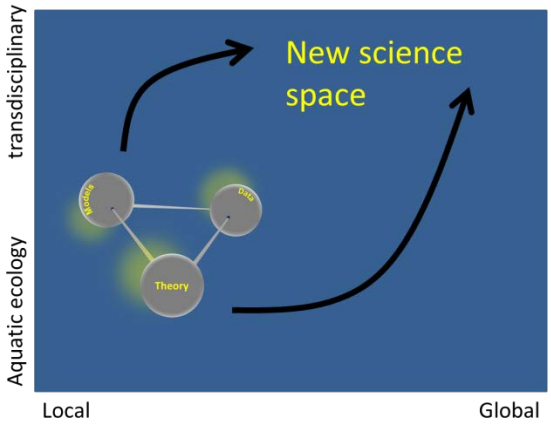

Figure 1. Theory, data and models are inextricably connected in science and supported by CI. GLEON is heading toward a new scientific space – one that is more global and transdisciplinary.

To enable the scientific expansion, GLEON has developed a number of technologies in an effort to create end-to-end solutions (Fig. 2). Data streaming, through DataTurbine, has enabled data flow to point-of-presence (POP) located at the field sites, as well as forwarding to a GLEON centralized database, Vega. Tools have been developed to download data from Vega and analyze in a variety of capacities, including in distributed computing infrastructure, such as Condor.

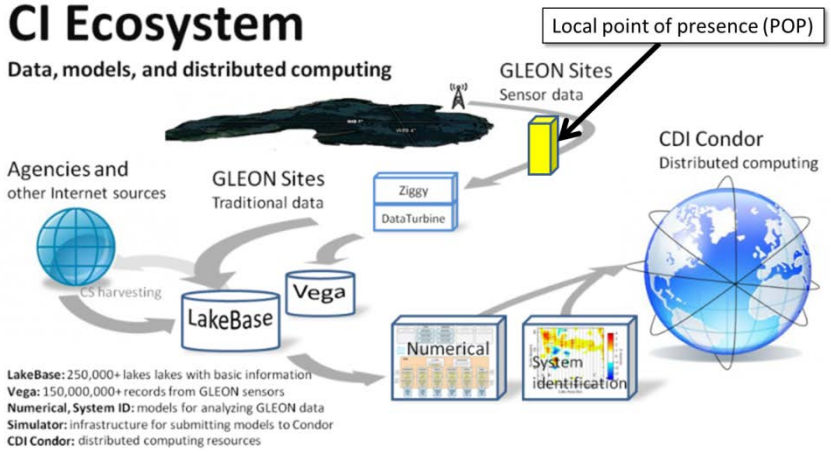

Figure 2. The end-to-end cyberinfrastructure (CI Ecosystem) of GLEON.

Through this expansion, GLEON has learned the value of CI tailored to the diverse needs of its members. However, GLEON's attempts have been met with mixed success. The successes include: site-level CI (from sensor to POP), in which GLEON sometimes plays a design and installation role; more sites coming online every year; visualization at the site level; proofing a technology (Vega, based on CUASHI's ODM) for storing large data sets; data analysis of large data sets; community development; partnering; science, which GLEON continues to produce. Unmet goals include: centralization – central repository has not really been used for science; persistence – little on-going contribution to the centralized system, even though sites continue to collect data; online data discovery – rarely used because data are not published in a consistent way; QA/QC at all levels; sharing – strings are attached; efficiency – the current system of exchanging data via email is very inefficient.

In this presentation, we hypothesize reasons for the successes and failure. These relate primarily to the culture of ecological science and how that is manifest in practice today. We highlight a few here:

- Data were collected for a specific experiment: Most ecological data are from field campaigns designed around specific questions and funded for the relative short-term. Thus, the original motivation was to service the immediate needs of the grant and not the data needs of the broader community. Restructuring data for discovery is significant effort.
- Incentives for anonymous sharing are not compelling to many: The probability of data collected from short-term field campaigns being discovered and used is relatively small today, even when the data are published online. Furthermore, attribution by data consumers, though likely, provides only modest value to the data provider in most cases.
- Online data discovery is not used much in ecology: There simply is not the critical mass of ecological data online and discoverable. Rather than use mixed modes of discovery (i.e., search online and email colleagues), scientists tend to resort to social network approaches, such as email.
- Ecologists have been slow to adopt standardization: Ecology is a diverse field made of many individual investigators who act independently. When data are not used outside the original project, there is little need for standardization. This may be a "chicken and egg" problem.
- General culture of "do it yourself": Ecologists are like entrepreneurs who feel they have to do it all – from personnel, to accounting, to science, and more.

Next, we postulate ways to reach the ecology of tomorrow.

- Address as cultural and technological set of issues that are linked
- New paradigm may require a new belief system and/or new accepted practice (more of a market system?)
- Use science projects to drive the CI (learning by doing), which has short-term payoff for both domain and CI scientists
- Determine what motivates ecologists to share data in more active, discoverable ways
    - Incentives need to match value system
    - Bar needs to be very low
    - We can learn from (or collaborate with!) others
- GLEON is entering a new era of science, in which ecologists are exploring new scales, technologies, new collaborations. Opportunities and innovations often happen at the seams between disciplines.
    - A new kind of ecology?
    - A new kind of ecologist?
    - PRAGMA Expedition

Tim **Kratz**, University of Wisconsin, *GLEON Background Presentation*

GLEON is an international grass-roots, collaborative network of limnologists, ecologists, information technology experts and engineers that collects and synthesizes high-frequency data from lakes worldwide to sense and forecast change. It was formed in 2004 and has grown to more than 300 individual members from six continents and more than 30 countries.

GLEON is an example of a new way of doing science. As a grassroots, member-driven organization, its members set the scientific direction, collect and analyze data, and advance scientific understanding of lakes. Over its history GLEON has evolved from making inferences at the scale of individual lakes to using data from multiple lakes in cross-site comparisons that allow for broader inference.

GLEON has identified a number of lessons learned including the following:

- Scientific networks are people networks
- Scientists around the world are eager to collaborate
- Students are central to vibrancy of the network
- Public is eager to engage with local science if it is placed in a global context
- Data sharing, both conceptually and technically, requires long-term sustained efforts

Chau Chin **Lin**, Taiwan Forest Research Institute, *Development of Ecoinformatics in EAP-ILTER*

Taiwan Ecological Research Network (TERN) was established 20 years ago in 1992. One of the major goals was to promote ecoinformatics. TERN started to collaborate with US LTER on ecoinformatics in 1995. Drs. TC Lin and ZG Pan were the two vanguards in early 1998 to the US-LTER headquarters learning IM techniques. In the same year, TERN hosted International LTER Global Communication Workshop which intended to develop a regional information management system. But there was little progress by funding deficiency. Not until an All Scientists Meeting was held in 2003 the TERN organized an IM team to US LTER sites looking for IM training in 2004. The ambition of building IM capacity was refueled and the TERN IM team members sought out collaboration with IM scientists in CAP, NTL, SEV and VCR LTER sites starting in 2004.

Fully supported by the NSF and NSC, a close IM training program between IM scientists from TERN and US-LTER (NTL and VCR) had been conducted from 2004 to 2006. The TERN has built a strong capacity to partner with the scientists of the EAP-ILTER network (such as the JaLTER, K-LTER, CERN, Philippines LTER and Australia-LTER) working regional data basically following the US-LTER IM system. And currently the EAP-ILTER network has closely linked to the ecoinforamtics groups in ILTER, PRIME, GBIF, GEOBON/APBON, DataOne communities.

The TERN IM Team will keep the step to seek collaboration with US LTER and regional partners to expand ecoinformatics into biodiversity informatics under the global trends of conservation and climate change research.

Activities History:
Many following activities promoted by the TERN were closely collaborated with US-LTER scientists

2004 Participating Lake Metabolism project (with NTL and UCSD)
2005 First EAP-ILTER I IM workshop in China (with US LTER and CERN)

2006 First US LTER-TERN ecoinforamtics workshop in Taiwan; Second EAP-ILTER IM workshop in Taiwan; EAP-ILTER IM committee Meeting in Japan; ILTER IM committee Meeting in Namibia; IM workshop in Philippines
2007 EAP-ILTER IM committee meeting in Taiwan; The third EAP-ILTER IM workshop in Korea ; Reciprocal visits between Australia LTER and TERN on ecoinforamtics issues; LTER IM workshop in Malaysia ; LTER IM workshop in Thailand ;
2008 First ILTER IM data sharing workshop in China; Participating DataOne project (TERN as one of nodes);
2009 Participation of Asia Pacific Biodiversity Observation System (APBON); First Forest Dynamic Plot information application workshop in Taiwan (with Malaysia LTER, JaLTER, US LTER, CTFS)
2010 Korea LTER IM workshop; The Second Forest Dynamic Plot Data Application workshop in Malaysia (with Malaysia LTER, Korea LTER, Singapore DP, Vietnam Biodiversity Center, US LTER); The Second US LTER-TERN IM workshop in Taiwan; Participation of the PRIME project (worked with 2 SCSD students)
2011 Participation of GBIF regional IM project; Participating GEOBON and APBON data management project (TERN will host data center in Taiwan); Participation of the PRIME project (worked with1 SDSC students);Participation of SensorPad project (with SDSC Super Computer Center); Finland LTER and TERN information system analysis project in Taiwan;
2012 Participation of PRIME project (will work with 2 SDSC students); GBIF EAP regional IPT2 workshop; GBIF EAP node manager meeting in Taiwan

Phil **Papadopoulos** and Peter **Arzberger**, University of California San Diego, *PRAGMA Overview and Directions*

PRAGMA, the Pacific Rim Application and Grid Middleware Assembly, established in 2002 through a series of workshops, enables small-to-medium size international groups to make rapid progress in conducting research and education by providing and developing international, experimental cyberinfrastructure. In addition, PRAGMA engages and builds long-term collaborations among its member institutions around the Pacific Rim. Currently there are more than 30 institutional members or participants in PRAGMA activities.

PRAGMA advances its mission through activities that

- Foster international “scientific expeditions” by forging teams of domain scientists and cyberinfrastructure researchers who develop and test necessary technologies to solve specific scientific questions and create usable, international-scale, cyber environments;
- Develop and improve a grassroots, international cyberinfrastructure for testing, computer science insight and advancing scientific applications by sharing resources, expertise, and software;
- Infuse new ideas from developing new researchers with experience in cross-border science to engaging strategic partners;
- Build and enhance the essential people-to-people trust and organization developed through regular face-to-face meetings, a core component of PRAGMA’s success.

Since 2002, PRAGMA has incubated new scientific networks like GLEON (Global Lakes Ecological Observation Network); played a critical role in the formation of programs that help train the next generation of cyber-scientists through research exchanges, e.g., PRIME (Pacific RIM undergraduate Experience) in the US, PRIUS (Japan), and MURPA (Australia); enabled the international membership of more than 30 institutions to respond to various disasters or emergencies including SARS and the Japanese Tsunami of 2011; created a persistent experimental test bed for international cyberinfrastructure; used the testbed to developed and tested software and advance science.

Some current expeditions and technology development projects include

- Predicting impact of eutrophication on lake ecosystem services; Understanding adaption in extreme ultramafic (serpentine) regions; Addressing infectious disease through computer aided drug discovery; Recovering IT services after severe disruptions
- Extending cyberinfrastructure to sensor-based observational networks; Building a PRAGMA Data Cloud through data sharing, provenance, data valuation and evolution experiments; Developing a trust envelop for collaboration through software-defined overlay networks and experiments with IPv6; Creating a multi-cloud environment through application-based migration experiments

Collectively we are working towards building a user-defined trusted environment of resources (compute, data, sensor or equipment) that will more rapidly enable collaborations among distributed teams of researchers.

Student participation is strongly encouraged, in particular in participating in expeditions and working in the PRAGMA Student group (motivated by the GLEON Graduate Student Association) to gain leadership experience.

For further information see http://www.pragma-grid.net/, and in particular our annual Collaborative Overviews: http://goc.pragma-grid.net/pragma-doc/overview/2012.pdf

Apivadee **Piyatumrong**, National Electronics and Computer Technology Center (NECTEC), *E-Rium: Environmental Informatorium*

E-Rium is a one-stop and multi-faceted observation data service. In particular, E-Rium enables users to simply (i) acquire any observation data of interest across geographically distributed repositories via a single point access; and (ii) visualize any observation data of choice in a desirable graphical presentation on demand and in a user-oriented manner.
Accordingly, it provides a one-stop service for data registry, discovery and access. However, it is likely that an observation data from different repositories can be heterogeneous in term of data structure and data coding scheme. Such heterogeneity strongly prohibits the interoperability of data across different repositories. Although, there exists metadata standard of some particular observation data, repository owners may or may not adopt the suitable metadata standard for their observation data. According to the previous stated fact, data integration feature in E-Rium is implemented at two different levels. Firstly, data transformation tool of E-Rium is one of the feature that, to the best of our knowledge, other portals do not provide yet. It basically transforms data from one to another format on-the-fly, given that the

format of both sides are generic formats (i.e. netCDF, .csv, .txt, etc.) This is a way to preserve the legacy systems of repositories. Moreover, with such tool, the interoperability across repositories can seamlessly be accomplished in a vertical manner. Secondly, data integration based on relation configuration among any related observation data is implemented. As an example, the interoperability of tidal and wind data facilitates data consumers to not only explore their correlation but also to have further decisions. The interoperability in this manner can be achieved by developing the relation configuration among any related metadata standards and horizontally integrating related observation data based on their relation configuration.
The applicability of E-Rium is currently evaluated by a few users. The result has shown that E-Rium enables users to simply acquire the preferable visualized observation data in a time and budget efficient manner. However, the result also show that further study must be done for finding more need of different users, especially from scientist user type. Last but not least, beside users, E-Rium encourages repository owners to register and share their belonging observation data.

Jordan **Read**, USGS, *USGS Activity*

The USGS has a long history of data federation and dissemination of these data to the public. Recent efforts have included the creation of web services for data access and visualization, as well as combined interagency efforts to centralize access points for large federal databases (such as the water quality portal; www.waterqualitydata.us). A group within the USGS, the Center for Integrated Data Analytics (CIDA) specializes in these types of science support tools. Much of CIDA's work is focused on the use and implementation of these tools that leverage data and data services that adhere to open standards (such as NetCDF, waterML2.0, and sensor observation services (SOS)). These tools are used by groups outside the USGS, and contribute towards the development of a larger, more integrated science network that spans agencies and user communities.

Kwang-Tsao **Shao**, Kun-Chi Lai, Yung-Chang Lin, Cheng-Hsin Hsu, Hsiang-Ying Li, Lee-Sea Chen, Guan-Shuo Mai, Han Lee. Biodiversity Research Center, Academia Sinica. *International Cooperation is the Key to the Success of Taiwan's Biodiversity Information Integration*

The biodiversity databases in Taiwan more than a decade ago were all dispersed, not integrated or linked to any international databases. The earliest international cooperation happened in 1994 when the Fish Database of Taiwan collaborated with the global FishBase. By joining Species 2000 and becoming GBIF's Associate Participant in 2001, Taiwan officially launched the work on international database cooperation. The first task is to establish and integrate Taiwan's own databases and websites. Fortunately, the Department of Humanities and Social Sciences of the National Science Council started funding a 2-phase 10-year "Taiwan e-Learning and Digital Archives Program" in 2002. The second phase of the project added the "International Collaboration & Promotion Project" in 2007. With the support of this project, all the animal and plant specimen collections in Taiwan are able to be digitized and integrated. In addition to specimen data, Biodiversity information includes species checklist, ecological distribution, species description, literature, audio-video, etc. Since scientific names are a primary key for

linking biodiversity databases, the species checklist database (COL) must be created first. To be built and linked afterwards are those of Barcode of Life (BOL), Encyclopedia of Life (EOL), and International Long Term Ecological Research (ILTER). This path is also used by the international community to develop biodiversity databases.

In 2004, the Biodiversity Research Center, Academia Sinica was formally established. Commissioned by National Science Council's Department of Life Sciences as well as Council of Agriculture's Forestry Bureau, the Center started in 2003, 2004, 2005, and 2012, respectively, to establish the databases of "Catalog of Life in Taiwan" (TaiBNET= TaiCOL), GBIF's Taiwan Portal (TaiBIF), Cryobanking of genetic material and DNA barcoding in Taiwan (TaiBOL), and Taiwan's Encyclopedia of Life (TaiEOL). These databases are linked to the corresponding international databases of COL, GBIF, BOL, and EOL. The data are exchanged and shared either through GBIF; or through international collaboration projects such as OBIS, GEO-BON, WoRMS, GenBank, IUCN-ISSG, and WDS; or through Global Species Databases of various organisms (GSD); or through individual country's database portal. So far, TaiBIF has aggregated 1.6 million specimen occurrence and observational records (1.2 million are geo-referenced point data), most of which have been provided to GBIF. By grasping the opportunities of international cooperation and adopting the mainstream of information technology over the past ten years, the Center not only has successfully integrated domestic cross-agency and cross-taxa biodiversity information, it has also effectively linked to the world biodiversity community—a model for database integration both in Taiwan and abroad. The main reasons for the success are: (1) the national policy, i.e. the requirements of Executive Yuan's "Biodiversity Promotion Plan"; (2) the aforementioned support of National Science Council and Council of Agriculture. Their funding provides adequate manpower and material for the tasks. Especially helpful is the budget allocated for attending international conferences and workshops; (3) the active participation in many international organizations and collaborative projects. It facilitates the learning and introducing of newly developed information technology and tools such as open source software to keep pace with the world's technological advances; (4) the educational training and promotion of the integration and publishing tools, including DiGIR, TAPIR, and IPT, to encourage the sharing of information; (5) the policies of sharing resources, technology, and research results, which include using CC licenses, offering technical support, and organizing training workshops; (6) the active promotion of top-down approach regarding information policy. The Executive Yuan's National Council for Sustainable Development requires governmental agencies to turn in specimen and ecological distribution raw data when a public-funded project ends. Due to the fact that Taiwan is not a member of the United Nations, it is very difficult to participate in international programs or be treated fairly. Nevertheless, Taiwan has a very rich biodiversity and enjoys working together with international partners to conserve biodiversity. The experience Taiwan has in successfully building biodiversity databases and collaborating with international counterparts can be shared with colleagues who are engaged in database integration.

Whey-Fone **Tsai** (NCHC), Bo Chen (NSPO), Chi **Wu** (TORI), Sornthep **Vannarat** (NECTEC), Shih-Ching **Lin** (NCHC), *Case Studies and Needs: Disaster Mitigation*

The objectives of Disaster Management Information Platform under NARL (National Applied Research Laboratories) in Taiwan is to establish the large data processing system for data backup & redundancy, circulation, sharing, model analyses, and visualization; to develop the system of communication, collaboration, and coordination utilized among the governmental disaster prevention and rescue units; and to accelerate the integration of multiple disaster data/information sources provided by the distributed government agencies and analyzed results to improve the performance of disaster response and mitigation operations. NCHC (National Center for High-performance Computing) is in charge of platform development. Disaster Management usually connects data from the multiple resources outside and only observes real-time, critical data which is small portion in total data. In the Disaster Management Information Platform project, the major data sources are from many government agencies, such as weather bureau, water resource agency, water & soil conservation bureau, and et al.; while NSPO (National Space Organization) and TORI (Taiwan Ocean Research Institute) are the major observed sources under NARL that contribute to this project. The data of observation used in the disaster management include satellite image, seabed bathymetry & geology, and ocean surface current. NSPO has contributed near real-time satellite images in many natural disasters for international humanity relief efforts. NSPO now is conducting joint operation development of multiple satellites for Disaster Service Management that will improve efficiency, lower cost, and obtain multi-scene data. TORI is also building its Ocean Database and Information Network to provide service to academia in Taiwan as well as international collaboration research, and conducting long-term Ocean Monitoring and Forecast research. For the natural disaster response in Thailand, NECTEC has started to make efforts on data collection/connecting and conduct research development on ocean simulation for storm surge and related development associated with flood disaster response. Under PRAGMA, NECTEC, AIST, and NARL are the key members in geosciences working group and disaster management collaboration.

In the Bridging Big Data Infrastructure workshop (Dec. 3-6, 2012), disaster management group (NARL, NECTEC), in addition to share the information indicated above, was able to interact with leaders of NSF Ecology/Environment/Biodiversity organization groups, such as GLEON, NEON, Biodiversity, DataOne, and EarthCube. Those organizations/groups have well developed in data framework, quality, provenance, and workflow. The sharing of the information in the workshop is very valuable that provides a vision of direction for disaster management's future development into maturity and completeness. In the PRAGMA platform, disaster management group can further cooperate with Ecology/Environment/Biodiversity organization groups in BIG DATA development and applications. Especially, disaster management group has better opportunity to get various disaster scenarios for best practice of BIG DATA, and is happy to share in PRAGMA and other associated activities.

Sornthep **Vannarat**, National Electronics and Computer Technology Center (NECTEC), *NECTEC and Disaster Mitigation Projects*

National Electronics and Computer Technology Center, NECTEC, is a government research center that develops technology solutions for Thailand's government and social sectors. NECTEC is a part of National Science and Technology Development Agency (NSTDA), Ministry of Science and Technology. NECTEC's primary mission is to help strengthen the sustainability of Thai industries and the sufficiency of Thai society through collaborative research and development for electronics and computer technologies. Some examples of NECTEC's projects include producing industrial-automation technology, assistive

technology for the disabled and elderly, and sensor network and computer modeling for environmental management.

Two projects of NECTEC that may benefit disaster mitigation are “Hydrodynamic Ocean Simulation” and “Hydraulic Modeling of Water Drainage in Urban and Industrial Area”. The first project has been conducted for three years. A program code called, Finite Volume Coastal Ocean Model, FVCOM, which is developed by UMASSD-WHOI joint efforts, is used to simulate the water level in Thai Gulf. Originally, the objective is to study the coastal erosion. However, it was shown that the model could accurately predict the effect of storm on the water level or storm surge. Therefore, it can also be used for disaster mitigation such as for predicting the extent of coastal flooding due to storm or predicting the seasonal water rise due to monsoon wind, which affects the water flow in rivers. NECTEC is working to extend the model to wider areas including the South China Sea and the Andaman Sea. The model will also be used to assist an ecological research of coral reef environment. This model uses the input data from GEBCO, OTIS, and Digital Typhoon project. The water level data recorded by Thai Navy is used to verify the model. In the future, NECTEC can share the simulated results.

Due to the flood in 2011, which covered a wide area of Thailand and caused a lot of damage and lost, NECTEC has started the “Hydraulic Modeling of Water Drainage in Urban and Industrial Area” project. The objective is to build the capacity to use numerical methods to assist flood prediction and flood disaster mitigation by study the drainage in a specific area. The area that is enclosed by a road system and has water channels to provide in-flow and out-flow has been chosen. This project was started eight months ago. NECTEC has installed some sensors to collect the water level data, and has collected rain fall data in the area. Afterward, the survey of land elevation and profiles of water channels will be conducted. The data will be used in a numerical model to simulate the water flow. When the project is completed, not only a numerical model will be developed, but the data collected will be useful for developing and testing other models.

Recently, there is a collaboration led by Electronics Government Agency, EGA, to develop a portal for data relevant to flood disaster mitigation. The participants are Geo-Informatics and Space Technology Development Agency, GISTDA, some NGOs and NECTEC. EGA has a mandate to develop IT infrastructure for government agencies and NGOs has experience in flood disaster mitigation. GISTDA is responsible for taking and managing satellite data. NECTEC has developed a platform for integrating data from various sources. The examples of data that this portal aims to provide are administration boundary, road, land use, places; citizen data, population density; hospitals, emergency medication services, refuge centers, volunteers; and water level, water channel, water gate, geographic data, flood report, flood warning and announcement. This collaboration is still in an early stage. The participants has begun to discuss the required data and work plan, and, started to discuss with data contributors.

Dave **Vieglais**, University of Kansas and DataONE, *About DataONE*

DataONE, the Data Observation Network for Earth (http://dataone.org), is a project sponsored by the US National Science Foundation in response to the DataNet solicitation, and has the general goals of enabling long term access to data of relevance to the environmental, ecological, and biodiversity sciences. To achieve this, cyberinfrastructure based on a common set of services interfaces and consisting of three major components has been deployed, and includes: 1) Member Nodes, which are data repositories exposing the DataONS service interfaces and operated by various organizations or groups interested in sharing their data; 2) Coordinating Nodes which maintain catalogs of data and metadata held on the Member Nodes, and provides

other services ensuring consistency of content; and 3) an Investigator Toolkit which is composed of various software tools, extensions, and plugins to enable access to all DataONE Member and Coordinating Node services in a consistent manner.

DataONE cyberinfrastructure provides a number of services that facilitate preservation, discovery, and reliable access to data. These services include provision of persistent unique identifiers for all content, replication of content across participating Member Nodes to ensure ongoing access in the face of node failure or deprecation, an identity mapping framework that helps ensure consistent application of access control rules across content regardless of the account a user has authenticated through, a centralized catalog and associated search index of all metadata to facilitate discovery, and emerging frameworks for provenance tracing and semantic integration of disparate datasets.

Brian **Wee**, NEON, *Overview of the National Ecological Observatory Network*

NEON is a user facility that provides free and openly available data and a variety of other resources for use by the public. Science user facilities provide resources that can be efficiently shared over many investigators and educators, or are too costly for individuals or institutions to maintain. Such facilities also make state-of-the-science capabilities accessible to a wide range of users. As ecologists begin to address ecology at the continental scale, the comprehensive data, spatial extent and remote sensing technology will allow a large and diverse user community to tackle new questions at scales not accessible to previous generations of ecologists. NEON's measurement strategy is designed to observe both the causes of ecological change (such as climate and land use) and biological responses to change from site to continent. NEON also coordinates operations with other research, networks, and observatories. All NEON data measurements are made consistently across NEON sites using rigorously managed calibrations and are traceable to national or internationally recognized standards.

F. List of Participants

Peter Arzberger, University of California San Diego
Reed Beaman, University of Florida
Cayelan Carey, University of Wisconsin
Bo Chen, Nation Space Program Office
Chih-Yu (Charles) Chiu, Academia Sinica
Hsiu-Mei Chou, National Center for High-Performance Computing (NCHC)
Chris Duffy, The Pennsylvania State University
Corinna Gries, University of Wisconsin:
Paul Hanson, University of Wisconsin,
Weicheng Huang, National Center for High-Performance Computing (NCHC)
Hen-Biau King, formerly Taiwan Forest Research Institute (TFRI)
Tim Kratz, University Wisconsin;
Jong Suk (Ruth) Lee, Korea Institute for Science and Technology Information (KISTI)
Hsiang-Ying (Clair) Li, Taiwan Biodiversity Information Facility (TaiBIF)
Chau-Chin Lin, Taiwan Forest Research Institute (TFRI)
Fang-Pang Lin, National Center for High-Performance Computing (NCHC)
Shyi-Ching Lin, National Center for High-Performance Computing (NCHC)
Sheng-Shan Lu, Taiwan Forest Research Institute (TFRI)
Guan-Shou (Jason) Mai, Taiwan Biodiversity Information Facility (TaiBIF)
Philip Papadopoulos, University of California San Diego (UCSD)
Aprivadee Piyatumrong, National Electronics and Computer Technology Center (NECTEC)
Beth Plale, Indiana University
Jordan Read, United States Geological Survey (USGS)
Kwang-Tsao Shao, Biodiversity Research Center, Academia Sincia
Steven Shiau, National Center for High-Performance Computing (NCHC)
Whey-Fone Tsai, National Center for High-Performance Computing (NCHC)
David Vieglas, University of Kansas
Sornthep Vannarat, National Electronics and Computer Technology Center (NECTEC)
Yu-Huang Wang, Taiwan Forest Research Institute (TFRI)
Brian Wee, National Ecological Observatory Network (NEON)
Jeng-Wei (David) Tsai, China Medical University, Taiwan
Uwe Woesner, University of Stuttgart
Chi Wu, Taiwan Ocean Research Institute (TORI)