The International Journal of Digital Curation
Issue 2, Volume 3 | 2008

Digital Data Practices and the Long Term Ecological Research Program Growing Global

Helena Karasti,
Department of Information Processing Science,
University of Oulu, Finland

Karen S. Baker,
Scripps Institution of Oceanography,
University of California, San Diego, USA

October 2008

Abstract

This paper explores data practices in a Long Term Ecological Research (LTER) setting. It describes a number of salient data characteristics that are specific to the LTER program and outlines some central features of the curation approach cultivated within the US LTER network. It goes on to identify recent developments within the international LTER program relating to data issues: increasing heterogeneities due to networking, integration of data from additional disciplines, and new technologies in a changing digital landscape. Information management experience within LTER provides one example of the recurrent balancing inherent to the work of data curation. It highlights (1) taking into account the extended temporal horizon of data care, (2) aligning support for data, science and information infrastructure, and (3) integrating site and network-level responsibilities. LTER contributes to the inquiry into how to manage the continuity of digital data and to our understanding of how to design a sustainable information infrastructure.

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. ISSN: 1746-8256 The IJDC is published by UKOLN at the University of Bath and is a publication of the Digital Curation Centre.

Introduction

This paper continues the line of work that describes and discusses data curation in specific disciplines (e.g. 1st International Digital Curation Conference[1]; 2nd International Digital Curation Conference[2]). We focus on a program anchored by field measurements in Long Term Ecological Research (LTER), representing a long-running network effort with long-term consistent data collection, preservation and access within the scientific community. LTER exemplifies a particular kind of setting for data practices, characterized and challenged by a long-term ecological science perspective coupled with an open data-sharing policy of primary research data in a highly distributed environment of interdisciplinary collaboration (Hobbie, Carpenter, Grimm, Gosz & Seastedt, 2003).

In effect, the US LTER program takes a "community approach" to data curation as distinguished from a research data collection or a reference data collection (National Science Board, 2005). More specifically, LTER takes a "site-based network approach" with data management at local levels as well as the Network Office. This places a portion of curation work close to the data sources, that is, hand in hand with ongoing scientific research. This approach ensures continuous data curation activities aimed at securing and providing access to "dynamic datasets" (Lord & Macdonald, 2003) within an extended temporal horizon. Balancing the needs of long-term data and ongoing science conduct, LTER information managers are an integral part of building information infrastructure for the network (ARL Workshop in Collaborative Relationships, 2006).

The aim in this paper is to provide a detailed description of data practices in LTER, in order to further understanding of what is at stake in actual data curation efforts of existing research collaborations. With data curation an emerging concept, we hope to contribute to its realization as a broad, integrative framework and as a complex role. The paper begins by describing the US LTER program and our study approach, followed by discussion of the most salient characteristics of LTER data and an elaboration on some elements central to the LTER approach to data curation. Finally, we point to recent developments in LTER scientific scope and global scaling, and also changing data practices. We conclude with some points on managing the continuity of digital data and building sustainable information infrastructure. Acronyms used in the paper are summarized in Table 1 below.

Studying the Long Term Ecological Research Program

The Long-Term Ecological Research (LTER) program was initiated in the United States by the National Science Foundation (NSF) in 1980 to augment the more typical ecological studies defined by short-term timeframes (Hobbie et al., 2003). The central organizing aim of the program, supported by long-term funding, is to understand long-term patterns and processes of ecological systems at multiple spatial and temporal scales. The US LTER network is challenged to foster this central aim while maintaining the diversity and independence of sites that comprise the network, i.e. to preserve simultaneously the quality of site science and the joint network activities. The program has grown gradually both in size within the United States and internationally

[1] 1st International Digital Curation Conference, September 2005 http://www.dcc.ac.uk/events/dcc-2005/
[2] 2nd International Digital Curation Conference November 2006 http://www.dcc.ac.uk/events/dcc-2006/

acronym	name	URL
US LTER	U.S. Long Term Ecological Research (LTER) Network	http://www.lternet.edu/
ILTER	International Long Term Ecological Research	http://www.ilternet.edu/
FinLTSER	Finnish Long-Term Socio-Ecological Research	http://www.environment.fi/syke/lter/
LTER-Europe	European Long-Term Ecosystem Research	http://www.lter-europe.ceh.ac.uk/
ALTER-Net	A Long-Term Biodiversity, Ecosystem and Awareness Research Network	http://www.alter-net.info/
SEEK	Science Environment for Ecological Knowledge	http://seek.ecoinformatics.org/
EML	Ecological Metadata Language	http://knb.ecoinformatics.org/software/metacat

Table 1. Acronym list with associated names and links.

(ILTER network) as well as in scientific scope (Haberl et al., 2006) to address global issues of great human concern. We use LTER to describe a global network of networks, representing existing and developing local, national, and regional sites, platforms, and networks.

Our interest in LTER lies particularly in its information management. Data issues have been on the US LTER agenda since the beginning of the network, and over the years their importance has increased, as reflected in the transition from the name of "data management" to that of "information management" (Baker et al., 2000). Extensive ethnographic fieldwork has been used for studying infrastructure (Star, 2002). A longitudinal study focusing on infrastructure was initiated in 2002 within the US LTER network and continues today; another longitudinal study within the Finnish Long-Term Socio-Ecological Research (FinLTSER) network commenced in 2007. A growing material corpus provides rich data for both individual and collaborative analyses. The quotations in this paper derive from interviews within the US LTER network, information managers are denoted by "(IM)" and scientists by "(S)".

Salient Characteristics of LTER Data

Ecological research frequently deals with extremely heterogeneous data: "We have a lot of varied types of datasets. Some studies may have huge volumes of records but not a lot of diversity, a 'deep database', like remote sensing. In ecological data in general you get much smaller datasets that cover a much wider variety, 'wide databases'. In general you are struggling with the diversity of different types of data. In genetics, for example, in comparison, databases are deep but not as complex." (IM) Further complexity of ecological datasets is introduced by missing values, midcourse modifications in sampling or procedures, addition or deletion of study parameters, plot or habitat modification by natural or anthropogenic disturbances or changing environmental conditions, and numerous other commonplace factors that lead to data anomalies (Michener, Brunt, Helly, Kirchner & Stafford, 1997). The ramifications of this complexity of the data are of great consequence to the conduct of science and of data curation.

LTER sites collect largely observational but also experimental data that contribute to an understanding of the local ecosystem as well as to development of central program themes. Some sites restrict data acquisition to datasets designated as "core"

for monitoring, whereas other sites preserve all data collected on their site's premises including short-term process studies. Such 'outdoor laboratories' of environmental field science have a history of manual data taking. Manual data taking allows data collectors to develop a close understanding of and relationship with the instrument as well as the ecological site. Arrangements are typically flexible allowing for emergent data gathering factors, such as in-the-field modifications. On one hand, this allows for science-in-the-making where analysis of and reflection upon the data begins already in the field or laboratory. On the other hand, such flexibility creates challenges for structured data flows and for updated or modified procedures while data collection continues. A summary of salient characteristics of LTER data activities is given in Table 2 below.

LTER dynamic data accrue seasonal additions and are subject to various kinds of revisions. Thus datasets require continuing curation. Quality control and analysis are performed together with updates to metadata. These elements of collecting, cleaning and preserving the data are part of the recurrent cycles of short-term local data use and publication. The long-term perspective further necessitates careful aligning of any new data accrued, assuring they "fit" with and continue an existing collection. This typically requires meticulous documentation of changes that have occurred.

The 'Long Term' aspect of LTER data introduces new possibilities of and requirements for reuse as new questions arise to be asked of existing datasets. Even "thoughts on why it's being collected and should it continue to be collected changing" (IM). In addition to the regular data collection, legacy datasets are recovered and curated retrospectively: "I was trying to document a lot of this historic stuff … I had a series of interviews with a PI who was coming on with Alzheimer's and I got incredible documentation for these early corporate data." (IM) Such efforts aimed at identifying and rescuing at-risk datasets not only prevent the loss of a site's longitudinal studies but also hold the potential of enhancing the development of a long-term perspective.

In the US LTER network the expectations for data access and delivery have evolved from well-curated data for site science purposes to open public access. Since the mid 1990's, NSF has aligned LTER funding with open access policies directed toward publicly funded scientific data (Arzberger et al., 2004; Porter & Callaghan, 1994), requiring sites to have primary research data available on the Internet two years after its collection. The new focus on sharing primary data online, rather than secondary/tertiary data or summaries in journal articles or reports, represents a significant expansion in the scope of responsibilities, moving from a need to understand materials within a scientist's career or within a project's timeframe to requirements for contextualization, preservation and access to primary data for wider-scale reuse over longer-term timeframes.

Long-term data defies the simplistic definition of "reuse" as "the use of data collected for one purpose to study a new problem" (Zimmerman, 2008). Rather, an individual long-term dataset can have multiple relationships with other datasets and with research questions during its lifetime. First, there is the "monitoring" aspect of data in which records are added periodically to a dataset. Immediate or short-term use of data yields a gradually developing, more informed understanding of the dataset as well as the local ecosystem. Analysis of annual additions in association with other

Activity	LTER characteristics
Data planning - local	Prospective in terms of fieldwork and local data use
	Site-specific ecological and social science data
	Largely non-reproducible observational but also experimental data
	Heterogeneous and complex data (sets)
	Largely manual data taking
Data acquisition	Attending to ongoing collection and updates over time to dynamic datasets
	Digital data record of experiments and process studies
	Retrospective recovery of legacy datasets in contemporary digital form
	Local data storage and preservation
Data description	Initial, intensive data & dataset description
	Continuing metadata description
	Multi-site data category building
	Local controlled vocabularies and dictionaries
Data use	Short-term analysis and use
	Site-based monitoring and innovative science
	Long-term network science
	Metadata update
Data delivery	Open, public access to data and metadata two years after collection
	Web interfaces for online data delivery
	Exchange with network partners and archives
Data reuse	Appropriate data presentation for direct use
	Appropriately contextualized delivery for data selection/integration
	Data preparation/structuring for data interoperability
	Unanticipated data uses
Data planning - global	Prospective in terms of data preservation
	Multi-community metadata standards making

Table 2. Salient characteristics of LTER data activities.

long-term datasets may lead to new hypotheses requiring more data gathering. or to immediately publishable results. Second, an individual LTER site provides a collaborative prompt to integrate across individual investigators' topics given some overarching shared themes such as local biome populations, state, and dynamics. The consequent sharing of data is a new use of the data. Third, investigators within the US LTER network engage in cross-biome, synthetic themes that rely on site data. Fourth, both within and outside LTER, data modelers are supported when long-term data are available. And finally, downloads of LTER data for reuse by non-LTER scientists, the general public and policy makers are another type of data reuse.

LTER data requires intensive description. First, there is the variability of ecological data. Such heterogeneities and complexities necessitate careful recording of contextual information starting at the time of data planning and subsequently during the actual data taking and data management. Second, in the context of open access, more data description is needed for (re)use situations distant from the origin of the data: "you have certain levels of metadata … if someone within the site was using the data, they know a lot about the whole collection system and the research system at the site, so you can give them less metadata … but to somebody outside or for somebody 20 or 30 years down the road, then it's going to be more and more critical that this whole story unfold." (IM)

While description of primary scientific data has always been integral to LTER data curation, the need for more standardized approaches has become increasingly crucial for data reuse. LTER sites have been faced with transforming tacit, informal

understandings of data nomenclature, methods, context, and quality into explicit procedures that can be incorporated into information management schemas. The LTER synthesis endeavors have brought to the fore what may be described as semantic and sociotechnical issues relating to exchange, integration, and interoperability. Along the temporal horizon of LTER data curation, work on these issues belongs to a prospective dimension, designing for the future.

LTER Approach to Data Curation

LTER data curation has an extended temporal horizon (Figure 1). The work carried out attends to 1) *ongoing* data collection and curation, 2) *retrospectively* recovering or tying to legacy datasets and 3) *prospectively* planning and designing enhanced possibilities for managing data (Karasti, Baker, & Halkola, 2006).

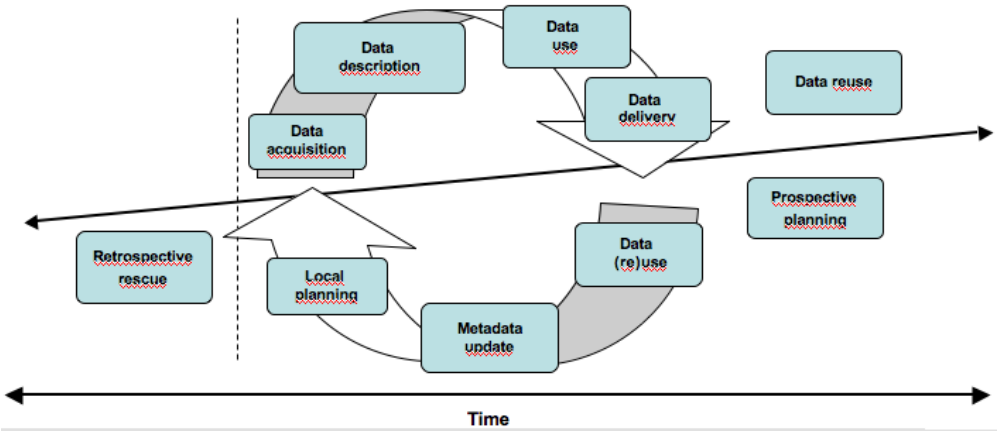

Figure 1. Extended temporal horizon of LTER data curation.

The predictable, recurrent elements (1) include the immediate-term issues of seasonal and annual cycles of data collection, entry to databases and preservation together with gathering the metadata; the near-term issues of data use and publication resulting from the two-year data policy stimulus for scientists to submit their data and metadata; and the long-term issues of data reuse and synthesis. The retrospective issues (2) of recovery of valuable datasets are not predictable but may, nevertheless, require rather urgent attention once recognized. Much of traditional ecology has been characterized by single investigator studies with strong personal ties to data (Zimmerman. 2008). Therefore an unknown number of datasets are still at risk of loss, even in the US LTER network that has been operating under a data-sharing model of science conduct. In addition, a complicating factor arises occasionally with the question of whether to tie a new dataset with an associated legacy dataset. With existing data practices, arrangements, and with available data curation resources limited, a continuous evaluation of priorities is required. The tendency is for more acute matters to take precedence. The prospective dimension (3) of data curation involves providing data for reuse, e.g. discovery and repurposing, but also requires consideration of a host of interrelated issues, such as protecting the longevity of legacy data, supporting ongoing research activities, keeping abreast with technology development, and looking for opportune funding openings. Thus, it is often impossible to coordinate with established or fixed time scales. For example, when and how to migrate elements of an information environment remains uncharted territory. Developing the skill for juggling arrangements to bring together the various temporal

dimensions into a working whole involves a great deal of tacit knowledge about data practices and data management priorities. This capability is an integral part of an information manager's expertise.

In US LTER, data curation is closely intertwined with information managers' other responsibilities relating to science and information infrastructure (Figure 2); the three activities together constitute what is called "information management" in the US LTER (Karasti & Baker, 2004). All US LTER sites carry out data curation because: "An unwritten rule is that … data are best managed at a site by people who know them … as far as quality control and assurance, and understanding the ways in which they were collected and the sites where they were collected." (IM) The value of local data curation performed by on-site information managers is that the understanding, engagement, and forward planning of information management can develop in conjunction with ecological field research and understandings, so that data directly and immediately enrich scientific investigations and site science provides focus for data curation. Thus, a close "two-way" relationship is formed between local data curation and science. Science support is not limited to provision of data and various kinds of assistance for collaborative science conduct. Information managers also promote data sharing and curation. Information managers have adopted a proactive attitude; they engage in motivating and educating scientists about open data sharing and long-term data management. "You need to convert them into thinking that putting data in our databank and on the web is something they really want to do. If they don't have the mindset that they want to share the data, it is really difficult to make them do it." (IM)

Data curation in US LTER is also closely intertwined with information infrastructure building based upon available technology. Historically, LTER has not been primarily driven by or focused on infrastructure development per se, rather the emphasis has been on developing technological support for the conduct of long-term science: "it's important that … information managers continue to come back to assessing whatever projects they want to develop to whether it is really going to support the research at the site." (IM). As technologies are developed at increasing speeds, staying technologically informed is an important aspect of an information manager's work: "It is a constant battle to stay current in technology." (IM). However, concern for the longevity of legacy data and the long-term research perspective underscore the merits of modest and unadventurous approaches to site information management. On one hand, incorporation of new capabilities to enhance data capture, use and preservation promises to facilitate science. Yet there is also present a concern for having in place a data-safe, functional system, "a protecting cocoon" (IM), for maintaining the integrity and availability of the long-term datasets. The features of high reliability, easy maintainability, and low risk for long-term data management influence judicious decisions about technology procurement. An information manager's foremost concern in aligning developing technologies with existing technologies and practices is to minimize disturbance of ongoing data preservation and use followed by interest in optimizing long-term data reuse and ease of maintenance: "The experience we have had with several of our things … the issue isn't how you do it, it's how do you maintain it and how do you make it so that it is easily maintainable." (IM)

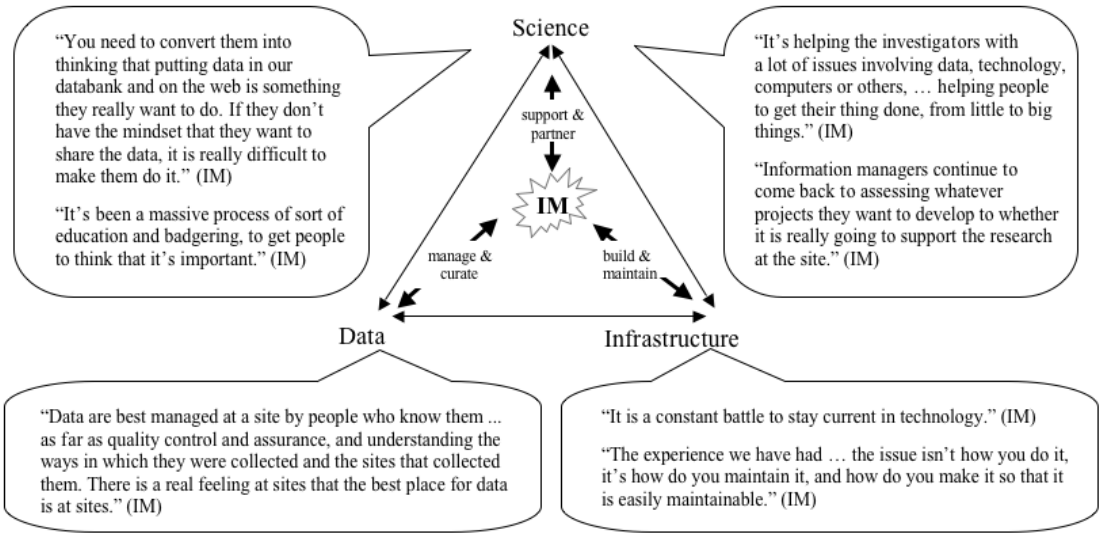

Figure 2. Three activities of LTER information managers: manage and curate **data**, support and partner with **science** as well as build and maintain information **infrastructure**.

Data curation is carried out both at the local or site level and at the network level, recognizing that each is committed to different responsibilities. This multilevel arrangement – a site-based network approach – was part of the initial organizational structure (Table 3).Despite their seemingly unadventurous, "feet-on-the-ground" approaches with technologies, information managers are proponents and play active roles in how technology and data management concepts are introduced and sustained at sites: "Researchers are looking at the information manager for guidance. Information managers need to be proactive and come up with their vision and plan for the site." (IM) Furthermore, "the information management community has been extremely proactive, and very responsive to demands at the network level" (S). Information managers see that their role is "really pivotal in leading the entire LTER community in recognizing the value of information technology and information management." (IM)

US LTER information managers have created a network level forum, an information management committee that forms a "Community of Practice" (Lave & Wenger, 1991; Wenger, 1998). Anchored by the realities and needs of their sites that reflect the histories and specificities of site science, information managers bring comprehension of the local settings to their network-level activities: "there are legitimate reasons for some differences between site systems" (IM) and appreciation for the heterogeneity of local infrastructures: "there is a variety of approaches among sites, and there is strength in diversity." (IM) Awareness of the long term provides an opportunity to develop a community with continuity that provides a reliable place for sharing and reciprocity: "It is good to see how other sites are doing things, either as a contrast or as an idea to improve." (IM) Information managers gain knowledge through working together: "LTER information managers have taken the time that fosters an integrative, sustainable approach with technology, ensuring that we learn together… It's like being mentored by the overall group." (IM) The network level community offers an arena for collaborative information infrastructure work. Information managers have created approaches and endogenous methods for jointly designing shared infrastructures that on the one hand rely upon inherent characteristics

of the networked organization, and on the other hand struggle with diversity and consensus (Karasti & Baker, 2004; Karasti, Baker & Halkola, 2006). The network level community gives the information managers a federated arena in which to develop specialist skills and professional identity in conjunction with the locally grounded work and routine day-to-day practices.

Initial LTER organizational concept	"And I think [names of two founding figures] could see that the long-term viability of the project really depended on being able to produce not just good site science but good network-level science. That was the way that the program had to justify itself in the eyes of Congress if it wanted to continue for many years of funding. … And frankly, again I think for understandable reasons there is a tension between trying to produce good publications from your site and manage your site, all those things the lead PI has to do. And then find time to also, to work with other sites. So I think it is true both on the scientific level and information management level." (IM)
Everyday practice of site-based network information management	"Where should the information managers' time be going? Should it be only to support site activities, or should some of it be going to support network activities? I have been interested in network-level activities and supporting them. … But I still have to be careful not to get over-involved. It is a balancing act, it always will be." (IM)

Table 3. As one aspect of their everyday practices, LTER information managers address multiple levels of community simultaneously, the local and the network.

To summarize, information managers in the US LTER provide support for rapidly developing technology, data requiring continuous care and science coping with data use in response to short-term evaluation cycles of scientific merit and long-term motive. Attending to all three - data care, science partnering and information infrastructure work - at site and network levels contributes to managing the longevity and continuity of the network's data and infrastructure.

Recent Developments in LTER Scientific Scope

Global issues, such as biodiversity, climate change, and ecological sustainability, give rise to scaling up of research to study complex issues of great worldwide importance. LTER collaboration has extended globally. Table 4 summarizes some of the developments in scope within the LTER. The US LTER network has grown from the initial six to current 26 sites. The International LTER (ILTER) network, founded in 1993 to develop a worldwide program and the infrastructure necessary to facilitate communication and information management (Gosz, 1999), now totals 38 national member networks. Regional networks have been developed with the European Long-term Ecosystem Research and Monitoring network (LTER-Europe), formed through the merger of the existing Western European with Central and Eastern European networks, as the most recent addition (Mirtl, 2007). New LTER sites, platforms, national, and regional networks are unevenly configured in terms of LTER science conduct and information management. Many of them have been operating within the traditional ecology research culture characterized by single investigator studies with strong ties to data and data sharing only between close associates (Zimmerman, 2008), and are challenged by a change in attitude towards open data sharing and large-scale collaboration. Awareness raising and education in data sharing and curation issues are needed among the scientists as well as information management personnel.

Furthermore, the spectrum of disciplines involved in LTER has broadened. During the first decade of the US LTER program, individual sites focused on their own long-term research projects. In the 1990s, spatial scales expanded and human-dominated ecosystems were incorporated in the form of urban-suburban sites. These extensions brought in first more varied ecological disciplines, followed by extension to social sciences. Within the ILTER network, the diversity of disciplines is even greater as the participants' programs do not necessarily follow the US model; some programs are much more structured and top-down with more emphasis on monitoring, whereas other programs have a greater regional focus and a stronger human dimension than most US LTER sites (Hobbie et al., 2003). For instance, the Finnish Long-Term Socio-Ecological Research network (FinLTSER), one of the most recent additions to the International and European networks, has a prominent emphasis on the "Social". Four out of the initial seven participants in FinLTSER network are LTSER areas or platforms having regional scope and explicit inclusion of socio-economic research; three are more traditional LTER sites. From the point of view of information management, the increasing diversity of disciplines creates more demands on data curation. Each new discipline that is added to the network brings locally specific terminology, adding to the integration challenges already posed.

In addition to a global pull, there is a technology push of cyberinfrastructure, e-Research, and e-infrastructure (National Science Foundation Cyberinfrastructure Council, 2007). New developments in sensor technology allow for accumulation of vast amounts of data, much greater than could ever be collected manually. New data encoding and transfer standards allow for transparent data representation to facilitate sharing of data from distributed sources. Increased bandwidth makes provision and access to vast datasets possible. These new data-related technology opportunities offer new directions to long-term ecological research that may profoundly change how science is conducted, and thus also require extensive changes in data curation approaches. All in all, the recent developments within the global LTER further increase heterogeneity in data, disciplines, institutions and technologies which, in turn, poses new challenges for LTER data curation and information infrastructure work.

'Global science pull'	Global issues give rise to • Broader scientific questions • Scaling up of collaborative research networks • Integrating human dimension (e.g. urban sites in USA and LTSER sites in Europe)
Spectrum of disciplines	Spectrum of disciplines involved has expanded • Ecological sciences and social sciences • Computer science and technological instrumentation • Information sciences, information systems and informatics
'Technology push'	Technology opportunities give rise to • Scaling up of digital networks • Automation of data generation and collection • Expansion of data scopes

Table 4. Recent developments in LTER scope of science, cross-discipline, and technology issues.

Global Science and Changing Data Practices

Changes in scientific scope together with new organizational arrangements for networking have ramifications for data practices and data curation. The wealth of additional data available from related disciplines and through international collaboration requires automated approaches to data discovery and to data use. However, this expansion in data scope brings with it a plethora of new terms and concepts introduced by the multilingualism entailed by the introduction of international datasets. Where the US LTER network is challenged by different terms referring to the same concept, i.e. Carbon Dioxide vs. CO_2, the international community must now also find mechanisms for integration across cultural and linguistic barriers. Coordination and negotiation on something as basic as a name, e.g. a species name, becomes an almost insurmountable task in the face of multiple, slightly differing species lists within each new country (cf. Bowker, 2000). Working with semantic issues, that is, issues of names and meanings, contributes to a cultural awareness of differences as well as advancing a common scientific landscape with significant policy consequences. For instance, there are policy ramifications dependent upon the concept of "sustainability". Yet the term sustainability has two quite different meanings: in ecology, it refers to assuring that available resources are not overused; in social science, it often refers to the economic viability of a process or a project over time.

The LTER networks provide examples of semantic approaches that have evolved over time. Current approaches to automated interfacing or reasoning with data draw on metadata associated with datasets. Metadata is the term used to describe data that provide local and general context to data collection programs, studies, datasets, and data columns. The US LTER Information Manager Committee adopted a non-geospatial metadata standard called the Ecological Metadata Language (EML) (Jones, Schildhauer, Reichman & Bowers, 2006; Michener et al., 1997) as a metadata specification for coordinating data access via a community catalogue. This standard is currently being established as part of a site's data delivery mechanism (Millerand & Bowker, in press). While this standard encompasses most topics required for scientific data reuse, it is semantically underdeveloped, allowing individual researchers to define their own terms and concepts for metadata annotation. There are currently working groups focusing on coordinating keywords in order to improve searchability of datasets and focusing on dictionary development as a community process rather than a static list. These are mechanisms for coordinating distributed development and providing continuity over time.

The ALTER-Net project, a Network of Excellence funded within the EU 6th Framework Programme, is at present designing a framework for data and knowledge sharing within the European LTER community (Schentz & Mirtl, 2003). After an initial evaluation of the EML standard defined by the US LTER, a more semantic approach was planned within Europe. A proto-ontological approach to data structuring and administration based on the data management system used within the Austrian LTER network has been proven to be effective for a wide range of LTER-related data. Consequently the ALTER-Net project is currently working on scaling this solution to the European level while migrating the technology to utilize the emerging Web Ontology Language (OWL). ALTER-Net has chosen a pragmatic approach to ontologies, leveraging the power of ontological approaches to data structuring, while

shying away from actual knowledge representation as this opens a prolific source of unforeseen troubles – a Pandora's box of approaches. The ALTER-Net ontology makes strong use of semantic relations for the description of concepts and their relationships to other concepts; it does not attempt to structure all required concepts within an ontological derivation hierarchy, as this is viewed as not economically feasible. Another concept anchoring this project is the seamless integration of metadata and data based on the philosophy that "one participant's data are another participant's metadata, and vice versa". (Schentz, Schleidt, Lane, Dirnböck & Peterseil, 2006.)

The work of ontology-building is a long-term effort being addressed in the US by the Science Environment and Ecological Knowledge project (SEEK), initiated as a NSF-sponsored effort designed to create cyberinfrastructure for ecological, environmental, and biodiversity research as well as to educate the ecological community about ecoinformatics. This project focuses both on the knowledge representation aspect of ontologies as well as the data structuring and metadata annotation capabilities as utilized within the scientific observation ontology OBOE.

As summarized in a US LTER review by information managers about multiple approaches to semantic issues, "semantic work challenges call for development of an assortment of strategies and collaborative mechanisms – all as part of a coordinated information infrastructure stretching from the immediate to the long-term" (Baker, Pennington & Porter, 2006). Initial methods of semantic clarification include the creation of lists of keywords and controlled vocabularies, relating entities within lists (dictionaries, thesauri and taxonomies), and developing conceptual relations (ontologies).

Ensuring multi-directional communications is necessary for effective infrastructure building. Technologists work with well-described data collections, encoding conceptual models in a computer-coded language. This work influences and should be influenced by the work of the scientific community. There is much new work to be done here: scientists are striving to articulate rapidly changing conceptual models influenced by insights gained from unprecedented availability of multiple types of information on varying spatial and temporal scales as well as from a growing diversity of cross-domain analyses; data and information managers are identifying relevant terms, potential categories, and use patterns.

Data work is further complicated by the continuing development of both knowledge representation and information infrastructure building as well as general technical advances. In the multiple interdependent arenas of science – individual labs, community repositories, regional or national archives, and networks of efforts – work with semantic issues is nascent. We are faced with improving interfaces with data through information systems while at the same time educating ourselves about knowledge representation and adding to our limited experience with semantic issues.

These multiple approaches to semantic issues address differing temporal spans of development. We are faced with handling simultaneously the different technical and social scales of semantic development, from long-term approaches (ontologies) to quicker methods (vocabularies and dictionaries) that inform ontology work. Human dimensions associated with the technical work include integrating semantic efforts and engagements across institutional, organizational, and cultural boundaries with their

differing legacy elements. These dimensions are intertwined with the great expectations placed in proposed solutions. An additional pair of concerns may be identified: there are more longer-term promises of progress than solutions, and the financial support required for the attainment of these solutions will either not be available or will be taken from existing research budgets.

Achievement of ambitious goals for digital data practices will only be possible through strong global cooperation and research into the actual practices of data-intensive scientific collaboration. All aspects of the described data process must be taken into account, starting with the digitization of legacy datasets and accrual of current data, spanning data storage, administration and quality assurance as well as metadata annotation and mapping of local concepts to global concepts in the form of standardized reference lists, thesauri and ontologies and including data discovery and integration tools. Only by working together, each element in the process providing a specific piece to this vast puzzle, will it be possible to create the cyberinfrastructure required to face the challenges posed to large-scale science today and in the future.

Concluding Words

The LTER program provides an experience-rich, thought-provoking example of data curation practices as well as an example of how networked data systems are developed. Managing continuity of data practices across sites as well as over the long-term is at the heart of LTER information management. This paper highlights certain key features of the data curation process undertaken by information managers situated locally at each LTER site. Features supported by a LTER shared vision include 1) a network-wide understanding of an extended temporal horizon; 2) recognition of managing as a whole the continuity of science, data and information infrastructure; 3) respect for the heterogeneities entailed through support of diverse approaches, taking into account multiple scales of development; and 4) a sustainable approach to infrastructure through an ecological, long-term site-based network model, in contrast to a limitless growth or non-sustainable 'endless frontier' model of building information infrastructure (cf. Bowker, Baker, Millerand & Ribes, in press). In addition, some features are supported by the LTER organizational structure itself: 1) the configuration of networking with different responsibilities at site, national, regional and international arenas, and 2) data management as part of the research environment with its ongoing commitment to support science in a bidirectional relationship, in contrast to idealized models that frequently portray data curation as taking place after data collection and submission have occurred.

A concerted effort is required to manage continuity and engage in data curation for new science needs while incorporating the sweep of new cyberinfrastructure, e-Research, and e-infrastructure development perspectives. With the LTER case study, there are two inseparable lines of practice: managing the continuity of digital data and growing a sustainable information infrastructure. LTER data curation requires a comprehensive vision, a continuous and longitudinal endeavor that must plan for change while assuring continuity. Furthermore, working with data requires making products available in the short-term while establishing processes for data work over the long-term. There is need for preserving, organizing, and providing access to long-term data in service and partnership with ongoing science conduct while building a sustainable information infrastructure able to handle change – whether it be change in resource, technique, or semantic structure.

The LTER program provides an example of a network facing new challenges as digital landscapes change. New instrumentation for data taking itself changes the way LTER science is conducted. With regard to increasingly automated data collection, there is the question of whether it represents a transition from or rather an augmentation of ongoing manual techniques. This cannot be predicted in advance and will require complex negotiations and alignments between related areas over extended periods of time to find suitable combinations, configurations and emphases. Two points emerge: 1) in addition to creating new practices, we should also study carefully what is at stake in existing practices (cf. Jirotka, Procter, Rodden, & Bowker, 2006), especially in fields with traditions of manual data taking, and 2) the change from "manual data taking" to "automated data life-cycle" represents a change in the diversity of data streams, requiring different and/or augmented data curation paradigms (cf. Lord & Macdonald, 2003).

As the LTER network grows, however, it also begins to face the challenge that "data curation needs to be addressed collaboratively at international level" (Beagrie, 2007). Emerging semantic frameworks offer some solutions to the problems posed by this new level of diversity, but in themselves also bring new challenges, both on a technological as well as on a social level. How to manage the needed communications in data curation and develop a global information infrastructure are critical questions that will depend upon the changing organizational, discipline, and career structures that emerge. New kinds of partnerships will be required to face this challenge ranging across technical, socio-technical, social, and domain-specific disciplines of LTSER.

The new digital landscape may be described as a diversity of "knowledge provinces" (Baker & Millerand, 2007) with related professional roles and skill sets (National Science Board, 2005) which illustrate the need to consider not only new types of careers but new types of education and learning environments. Broad vision and sensitivity to change will be required for managing continuity in the changing digital landscape of increasingly global collaboration and associated increasingly diversified data, discipline, research, and information management. We must not stop short of asking: how do we envision the sustainable international network of scientific networks supported by information infrastructures that include comprehensive data curation? In this we see the importance of cultivating new types of collaboration and cross-fertilization, that in addition to digital data preservation and curation (e.g. European Task Force Permanent Access, 2005), cyberinfrastructure/e-Research include several research fields such as Informatics, Social Informatics and Computer Supported Cooperative Work (e.g. Jirotka et al., 2006), Infrastructure Studies (e.g. Edwards, Jackson, Bowker & Knobel, 2007), and Social Studies of Science and Technology (e.g. Bowker, 2000).

There is a vast wealth of data and information available. However, it is up to us to recognize and configure resources for both data and technology as well as multiple levels of expertise and experience available within scientific site-based networks in order to provide an infrastructure for the facilitation of timely responses to tomorrow's questions based on yesterday's, today's and tomorrow's data.

Acknowledgements

This paper is an extended version of Karasti, H., Baker K.S., & Schleidt, K.: Digital Data Practices and the Long Term Ecological Research Program presented in

the 3rd International Digital Curation Conference, December 11-13 2007, Washington DC, USA that won the best peer-reviewed paper award of the conference. We thank Katharina Schleidt for contributions to the earlier version. We offer special thanks to the participants of the LT(S)ER networks with whom we have studied and collaborated. Support has been provided by Academy of Finland #119814 and #125467 (H. Karasti), by NSF OCE #04-05069, OPP #02-17282, SBE/SES #04-33369 and SIO, Ocean Informatics (K. Baker).

References

ARL Workshop in New Collaborative Relationships. (2006). *To stand the test of time: Long-term stewardship of digital data sets in science and engineering*. A Report to the National Science Foundation from the ARL Workshop on New Collaborative Relationships: The Role of Academic Libraries in the Digital Data Universe. Arlington, VA.

Arzberger, P., Schroeder, P., Beaulieu, A., Bowker, G., Casey, K., Laaksonen, L., et al. (2004). An international framework to promote access to data. *Science 303*(5665 (March 19)), pp. 1777-1778.

Baker, K.S., Benson, B.J., Henshaw, D.L., Blodgett, D., Porter, J.H., & Stafford S.G. (2000). Evolution of a multisite network information system: The LTER information management paradigm. *BioScience 50*(11), pp. 963–978.

Baker, K. S., & Millerand, F. (2007). Scientific information infrastructure design: Information environments and knowledge provinces. In *Proceedings of the American Society for Information Science and Technology*, (ASIST 2007).

Baker, K. S., Pennington, D., & Porter, J. (2006, Spring). Multiple approaches to semantic issues: Vocabularies, dictionaries and ontologies. *LTER DataBits Newsletter*. Retrieved November 22, 2008, from http://intranet.lternet.edu/archives/documents/Newsletters/DataBits/06spring/#3fa

Beagrie, N. (2007). *e-Infrastructure strategy for research*. Final report from the OSI Preservation and Curation Working Group.

Bowker, G. C. (2000). Biodiversity datadiversity. *Social Studies of Science 30*(5) pp. 643–683.

Bowker, G. C., Baker, K. S., Millerand, F., and Ribes, D. (in press). Towards information infrastructure studies: Ways of knowing in a networked environment. In J. Hunsinger, M. Allen & L. Klasrup (Eds.), *International handbook of Internet research*. Springer.

Edwards, P.N., Jackson, S.J., Bowker, G.C., & Knobel, C.P. (2007). *Understanding infrastructure: Dynamics, tensions, and design*. NSF Report of a Workshop on History & Theory of Infrastructure: Lessons for New Scientific

Cyberinfrastructures: 50. Retrieved November 22, 2008, from http://hdl.handle.net/2027.42/49353

European Task Force Permanent Access (2005). *Permanent access to the records of science: Strategic action programme 2006-2010*. The Hague: National Library of the Netherlands.

Gosz, J. R. (1999). *International long term ecological research: collaboration among national networks of research sites for a global understanding*. ILTER Regional Workshop, September 16-18, 1998. Madrain, Poland.

Haberl, H., Winiwarter, V., Andersson, K., Ayres, R.U., Boone, C., Castillo, A., et al. (2006). From LTER to LTSER: Conceptualizing the socioeconomic dimension of long-term socioecological research. *Ecology and Society 11*(2) p.13.

Hobbie, J.E., Carpenter, S.R., Grimm, N.B., Gosz, J.R., & Seastedt, T.R. (2003). The US long term ecological research program. *BioScience 53*(1) pp. 21–32.

Jirotka, M., Procter, R., Rodden, T., & Bowker, G.. (2006). Special issue: Collaboration in e-research. *Computer Supported Cooperative Work 15*(4), pp. 251-255.

Jones, M.B., Schildhauer, M.P., Reichman, O.J., & Bowers, S. (2006). The new bioinformatics: Integrating ecological data from the gene to the biosphere. *Annual Review of Ecology, Evolution, and Systemics 37*, pp. 519-544.

Karasti, H., & Baker, K. (2004). Infrastructuring for the long-term: ecological information management. *Hawaii International Conference on System Sciences 2004 (HICSS'37)*, January 5-8, 2004, Hawaii, USA.

Karasti, H., Baker, K.S., & Halkola, E. (2006). "Enriching the Notion of Data Curation in E-Science: Data Managing and Information Infrastructuring in the Long Term Ecological Research (LTER) Network." Computer Supported Cooperative Work 15(4): 321-358.

Lave, J., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation*. Cambridge, UK: Cambridge University Press.

Lord, P., & Macdonald, A. (2003). *e-Science curation report-Data curation for e-science in the UK: An audit to establish requirements for future curation and provision*. Twickenham, UK: The Digital Archiving Consultancy Limited.

Michener, W.K., Brunt, J.W., Helly, J.J., Kirchner, T.B., & Stafford, S.G.. (1997). Nongeospatial metadata for the ecological sciences. *Ecological Applications 7*(1), pp. 330– 342.

Millerand, F., & Bowker, G. C. (in press). Metadata standard: trajectories and enactment in the life of an ontology. In M. Lampland & S. L. Star (Eds.), *Formalizing practices: reckoning with standards, numbers and models in science and everyday life*., New York: Cornell University Press.

Mirtl, M. (2007, January). The big picture - Long-term socio-ecological research (LTSER) in Europe. *ALTERNews*, *03*. Retrieved November 22, 2008, from http://www.alter-net.info/SITE/UPLOAD/DOCUMENT/News/2007/Alternews_03_web.pdf

National Science Board (2005). *Long lived digital data collections: Enabling research and education in the 21st century*. National Science Board (NSB-05-40, May 23, 2005).

National Science Foundation Cyberinfrastructure Council. (2007). *Cyberinfrastructure vision for 21st century discovery*.

Porter, J. H., & Callahan, J. T. (1994). Circumventing a dilemma: Historical approaches to data sharing in ecological research. In W. K. Michener, J. W. Brunt & S. G. Stafford (Eds.), *Environmental information management and analysis: Ecosystem to global scales* (pp. 193-202).

Schentz, H., & Mirtl, M. (2003). MORIS an universal information system for environmental monitoring. Environmental Software Systems. In G. P. Schimak, D. A. Swayne, N. W. T. Quinn, & R. Denzer (Eds.), *IFIP Conference Series 5*, pp. 60-68.

Schentz, H., Schleidt, K., Lane, M., Dirnböck, T. & Peterseil, J. (2006). Functional concept for a biodiversity and ecological network, A long-term biodiversity, ecosystem and awareness research network. (ALTER-Net) Project no. GOCE-CT-2003-505298. WPI6_2006_04. p. 75. Retrieved November 22, 2008, from http://www.alter-net.info/SITE/UPLOAD/DOCUMENT/outputs%5CANet_WPI6_2006_04_Network_Concept.pdf

Star, S. L. (2002). Infrastructure and ethnographic practice: Working on the fringes. *Scandinavian Journal of Information Systems, 14*(2), pp. 107-122.

Wenger, E. (1998). *Communities of practice: Learning, meaning and identity*. Cambridge, UK: Cambridge University Press.

Zimmerman, A.S. (2008). New knowledge from old data: The role of standards in the sharing and reuse of ecological data. *Science, Technology & Human Values, 33*(5), pp. 631-652.