Ecological Informatics xxx (2016) xxx–xxx

Contents lists available at ScienceDirect

Ecological Informatics

journal homepage: www.elsevier.com/locate/ecolinf

Information management at the North Temperate Lakes Long-term Ecological Research site — Successful support of research in a large, diverse, and long running project

Corinna Gries *, Mark R. Gahler, Paul C. Hanson, Timothy K. Kratz [1], Emily H. Stanley

Center for Limnology, University of Wisconsin, 680 North Park Street, Madison, WI 53706, USA

ARTICLE INFO

Article history:
Received 18 February 2016
Received in revised form 22 August 2016
Accepted 28 August 2016
Available online xxxx

Keywords:
Ecological information management
Long-term ecological research
Database
Sensor data
Data life cycle

ABSTRACT

Information management has been an integral part of the research process at the North Temperate Lakes Long-term Ecological Research (NTL LTER) program for over 30 years. A combination of factors has made the information management system (IMS) at NTL very successful. Significant resources have been invested in the IMS from the beginning, the Information Manager has been part of the leadership team at NTL and later in various roles at the LTER network level; the NTL IMS was a very early adopter of database systems, standardized metadata, and a data delivery system based on those metadata. This approach has made data easily accessible to NTL researchers and the broader scientific community. Data management workflows have become increasingly more automated with adoption of modern technologies as they became available, making the system efficient enough to handle core data as well as all one-time research data generated within NTL and several related projects. More than three decades of core data from eleven lakes are reused extensively as critical background information and as the limnological go-to site for many synthesis projects within and beyond LTER.

The NTL IMS continues to implement new technologies for improving data management efficiency, discovery, access, integration, and synthesis. Accordingly, the functionality of the original online data access system programmed in Java and JavaServer Pages (JSP) was ported to the modern content management system, Drupal and integrated into LTER's Drupal Ecological Information Management System (DEIMS). NTL has invested in sensor technology for studying lake conditions over the long term, which necessitated a sophisticated management system tailored to high frequency data streams. Several technologies have been used at different times for automation of management, quality control and archiving of these high volume data. Near real time lake conditions can be accessed on the NTL website and smart phone Apps.

Easy access to long-term and sensor data in the NTL IMS has led NTL researchers to develop new analytical methods and the publication of several R statistical packages. Recent graduate students are now employed as data scientists helping define a new career path inspired by the availability of data.

The NTL project has amassed one of the world's most comprehensive long-term datasets on lakes and their surrounding landscapes. The NTL IMS facilitates the use of these data by multiple groups for research, education, and communication of science to the public.

1. Introduction to NTL

The North Temperate Lakes Long-term Ecological Research (NTL[2] LTER[3] https://lter.limnology.wisc.edu) program began in 1981 as one of six sites that formed the initial nodes of the LTER network funded by the National Science Foundation (NSF). Since then the network has grown to its current configuration of 24 sites conducting long-term research across a wide range of ecosystems (Waide and Thomas, 2012). In its fourth decade and located in the state of Wisconsin, USA (Fig. 1), NTL LTER studies the dynamics of temperate lakes in a landscape context and over time scales ranging from minutes, hours, days, and weeks to many decades and even millennia (Magnuson et al., 2006). It is guided by the overarching research question, 'How do biophysical setting, climate, and land use and cover interact to shape lake dynamics and organization in the past, present and future?' Throughout its 30+ year history, the NTL LTER program has made significant progress in understanding long-term change in two suites of lakes in the Northern Highland Lake District (NHLD; northern Wisconsin) and the Yahara Lake District (YLD, southern Wisconsin) (Fig. 1) along with their surrounding landscapes.

* Corresponding author.
E-mail addresses: cgries@wisc.edu (C. Gries), gahler@wisc.edu (M.R. Gahler), pchanson@wisc.edu (P.C. Hanson), tkkratz@wisc.edu, tkratz@nsf.gov (T.K. Kratz), ehstanley@wisc.edu (E.H. Stanley).
[1] Currently: National Science Foundation, 4201 Wilson Boulevard, Arlington, Virginia 22230, USA.
[2] NTL: North Temperate Lakes.
[3] LTER: Long-term Ecological Research Program.

http://dx.doi.org/10.1016/j.ecoinf.2016.08.007
1574-9541/

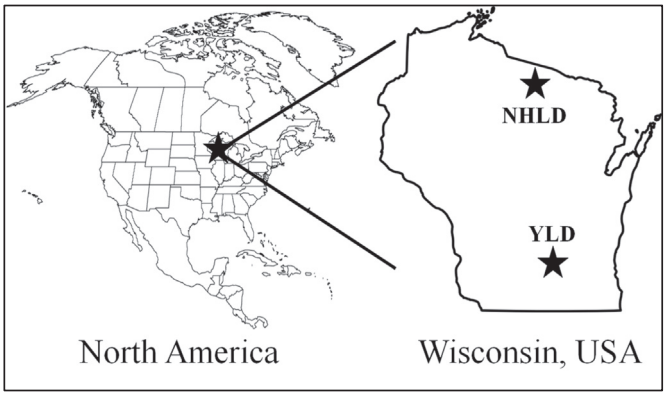

Fig. 1. Map of NTL locations within North America and Wisconsin.

At the NTL LTER program, Information Management (IM[4]) has been integrated into the research process from the very beginning of the project. In support of long-term research, the curation, storage, and reuse of data, i.e., IM, has been held to high rigor and standards that underlie the larger scientific effort. Although practices have changed throughout the past 30 + years, NTL IM has been guided by the goals of: (1) facilitating the scientific process; (2) applying appropriate technologies to enhance efficiency in data curation; (3) developing and following best practices for archiving and making data discoverable for future science; (4) actively seeking outside connections; and (5) fostering an environment of experimentation and innovation. The IM team embraces the full data life cycle and is actively involved in promoting and supporting Open Science (Hampton et al., 2015) through providing appropriate technologies and training.

2. History of information management at NTL

NTL has collected, curated, and made publicly available over 250 data sets that include physical, chemical, biological, and socio-economic variables for the eleven main study lakes and their watersheds (Fig. 2). These data are among the most complete long-term, multi-lake datasets in the world and provide the foundation for understanding long-term change in lakes and their surrounding landscapes. Included are core datasets — many of which began with project inception in 1981, results of short term investigations (most commonly from graduate research), manipulative experiments, larger regional studies, and some datasets that extend the long-term record by digitizing archived information as far back as the early 1900s.

Three decades of IM have seen many changes in technology as well as a major shift in the perception of why information management and open data exchange are valuable. Long-term core datasets have always been available to all participants for primary research to provide critical background information and for scaling of short-term research results in time and space at NTL. Hence, the culture of data sharing was established from the beginning, with the Information Management System (IMS[5]) striving to make data access as convenient as possible. In contrast, data from short-term research projects have taken longer to fully capture and publish within the IMS. However, due to a change in research and data management culture, the habit of submitting data to the IMS is becoming more widespread. A major driver of this shift is the requirement by a growing number of journals to have data published in a stable repository before the research paper is published.

NTL embraced database technology early in its development, with data being stored in a hierarchical database on the campus mainframe as early as 1982 (Benson et al., 2006). During the first decade, NTL added data loggers, a local area network, and a bibliography database before moving the main database into the fully relational Oracle database system in 1993. The next decade was marked by the rise of internet technology, and the first website, published in 1994, provided access to data and the site's bibliography. Researchers were trained in querying the central database using a commercial client application before sophisticated data querying and download capability was added to the website in 1997 for climate datasets. Oracle was also used to store metadata — that is, data about the data. This allowed the rapid expansion of initial querying capabilities to all datasets stored in Oracle and added download capabilities for datasets generated in proprietary formats, specifically, from Geographic Information Systems and molecular analyses. A Java/JavaServer Pages web application based on these metadata replaced the desktop client application in 2001 and provided data access well beyond the participants of the NTL project. The principles of using standardized metadata held in a relational database to build a web application are still being used (see below). These highly structured metadata strongly contrasted the previous unstructured text documents describing the data, and their development proved to be a great foresight when the metadata standard, the Ecological Metadata Language (EML[6] (Fegraus et al., 2005)) was adopted by the LTER network as a whole in 2002. Conforming to this new standard proved to be relatively easy and was accomplished by adding a few fields in the database and then exporting the data programmatically into the EML format — again, a procedure that is still implemented.

Throughout the history of NTL, the IM team has sought to improve efficiency through automation of quality control and data ingestion into the IMS. Given established workflows of the field and lab crews, it turned out that for most cases standardized Excel spreadsheets were the most effective means of transferring data to the IM team. Those data were then ingested into the database via Perl scripts and quality controlled with database triggers in Oracle. However, a few specific data entry applications have been developed which support better data quality, speed of sampling, or management of samples and results (MobileFish, Z3, and ChemLab; see below).

3. Information management personnel

Looking back over the 35-year history of the project, it is clear that the NTL IM effort has scaled with the scientific effort. It has been paramount that the skills and background of the information manager overlap those of the rest of the NTL team to provide for the effective integration of IM into the project. Over the years, members of the IM team have changed, but the team has always been led by an ecologist knowledgeable in information management who was able to communicate with ecologists, computer scientists, and informaticians. With minimal turnover, this person has provided leadership in developing an IMS that has served the site researchers well and successfully migrated it through emerging technologies. Other members of the team generally were more technically oriented and provided programming and systems administration skills. Student programmers were hired for specific projects.

This approach parallels that of the science and funding in NTL, with the vision for long-term data collection developed by more senior personnel being supplemented with faster turnover by graduate students. On the IM side, long-term visioning and continuity provided by senior IM staff is supplemented by short-term contributions by undergraduate computer science students. This also reflects the realities of funding, with long-term continuity provided by LTER, and shorter-term projects provided by related grants.

4. The current information management system

The majority of data life cycles at NTL start in the field. A field crew is responsible for core sample collection and processing and eventually

[4] IM: Information Management.
[5] IMS: Information Management System.
[6] EML: Ecological Metadata Language; https://knb.ecoinformatics.org/#external//emlparser/docs/index.html.

Please cite this article as: Gries, C., et al., Information management at the North Temperate Lakes Long-term Ecological Research site — Successful support of research in a lar..., Ecological Informatics (2016), http://dx.doi.org/10.1016/j.ecoinf.2016.08.007