Ecological Informatics 29 (2015) 33–44

Contents lists available at ScienceDirect

Ecological Informatics

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

Ecological data sharing

William K. Michener *

College of University Libraries & Learning Sciences, MSC04 2815, University of New Mexico, NM 87131-0001, United States

ARTICLE INFO

Article history:
Received 4 March 2015
Received in revised form 18 June 2015
Accepted 24 June 2015
Available online 2 July 2015

Keywords:
Data publication
Data sharing
Information technology
Metadata
Open access
Policy

ABSTRACT

Data sharing is the practice of making data available for use by others. Ecologists are increasingly generating and sharing an immense volume of data. Such data may serve to augment existing data collections and can be used for synthesis efforts such as meta-analysis, for parameterizing models, and for verifying research results (i.e., study reproducibility). Large volumes of ecological data may be readily available through institutions or data repositories that are the most comprehensive available and can serve as the core of ecological analysis. Ecological data are also employed outside the research context and are used for decision-making, natural resource management, education, and other purposes. Data sharing has a long history in many domains such as oceanography and the biodiversity sciences (e.g., taxonomic data and museum specimens), but has emerged relatively recently in the ecological sciences.

A review of several of the large international and national ecological research programs that have emerged since the mid-1900s highlights the initial failures and more recent successes as well as the underlying causes—from a near absence of effective policies to the emergence of community and data sharing policies coupled with the development and adoption of data and metadata standards and enabling tools. Sociocultural change and the move towards more open science have evolved more rapidly over the past two decades in response to new requirements set forth by governmental organizations, publishers and professional societies. As the scientific culture has changed so has the cyberinfrastructure landscape. The introduction of community-based data repositories, data and metadata standards, software tools, persistent identifiers, and federated search and discovery have all helped promulgate data sharing. Nevertheless, there are many challenges and opportunities especially as we move towards more open science. Cyberinfrastructure challenges include a paucity of easy-to-use metadata management systems, significant difficulties in assessing data quality and provenance, and an absence of analytical and visualization approaches that facilitate data integration and harmonization. Challenges and opportunities abound in the sociocultural arena where funders, researchers, and publishers all have a stake in clarifying policies, roles and responsibilities, as well as in incentivizing data sharing. A set of best practices and examples of software tools are presented that can enable research transparency, reproducibility and new knowledge by facilitating idea generation, research planning, data management and the dissemination of data and results.

Contents

* Tel.: +1 505 220 3123; fax: +1 505 277 2541.
E-mail address: william.michener@gmail.com.

http://dx.doi.org/10.1016/j.ecoinf.2015.06.010

1. Introduction

Data sharing is the practice of making data available for use by others. Ecologists are increasingly generating and sharing immense amounts of data as part of the research enterprise. The data are derived from direct human observations in the field and recorded in notebooks and other media, laboratory observations, remote and in situ sensors, and instruments that are employed to measure particular attributes of biota (e.g., presence, temperature) and the physical environment (e.g., air, soil, water) such as rainfall, solar radiation, soil moisture, and pH. Ecologists often use shared data that originate from other scientists for comparative purposes or to augment their data collections, for synthesis efforts such as meta-analysis, for parameterizing models, and for verification of results (i.e., study reproducibility). In some cases, shared data may be the only data or the best data that are readily available. Data are also used outside the research context. Many non-researchers use available data for decision-making, natural resource management, education, and other purposes.

Some science domains such as oceanography and taxonomy have a relatively long tradition of data sharing. For example, the International Oceanographic Data and Information Exchange of the Intergovernmental Oceanographic Commission (IOC) of UNESCO was established in 1961 to facilitate the international exchange of oceanographic data and information exchange (http://www.iode.org). The IOC has enabled the creation of more than 80 oceanographic data centers in IOC countries.

Data sharing in ecology, on the other hand, has evolved slowly and is only now becoming common practice. In this paper, I first describe the history of data sharing in ecology, primarily focusing on several of the large international and national (primarily USA) ecological research programs that have emerged since the mid-1900s. Second, I examine the sociological aspects of data sharing, especially the perceived impediments and benefits, and review the role of societies, funders, and journals in changing the culture of data sharing. Third, I review the role of cyberinfrastructure in supporting data sharing including data repositories, software tools, persistent identifiers, and federated search and discovery. Last, I discuss the future of data sharing and conclude with a set of best practices for sharing ecological data.

2. Ecological data and a brief history of data sharing

In a review of historic ecological data, Bowser (1994) categorized ecological data into three types: (1) planned—i.e., well-planned and well-documented long-term data such as the long-term records of atmospheric CO_2 from Mauna Loa, Hawaii (Keeling et al., 1976) and the Hubbard Brook watershed studies in New Hampshire (Likens et al., 1977) that were relatively rare at the time; (2) opportunistic—i.e., data that are collected to achieve short-term goals over a discrete funding period and are commonly encountered in the literature; and (3) serendipitous—i.e., data that are not for testing a scientific hypothesis such as weather data collected by private citizens, fish and wildlife harvest data, and other types of data. Bowser (1994) recounted efforts that began in 1979 at the North Temperate Lakes Long-Term Ecological Research site to retrieve and use data previously collected in Wisconsin lakes including the data sets generated in the pioneering limnology studies by Birge and Juday that led to more than 400 publications over a period of seven decades (see Juday and Hasler, 1946). Bowser (1994) summarized the state of the historic data as:

> "*The scope, degree of documentation, quality, and availability of different data sets varies widely. Both published and unpublished data sets have strengths and weaknesses. Data discontinuity, whether from single or multiple sources, makes data calibration difficult. Quality control is uneven, at best, and is often undocumented. Instrumentation changes have been rapid and intercalibration with new techniques is not practiced as commonly as would be hoped.*"

Such data challenges are not unexpected in an emerging, but relatively young scientific discipline. Prior to and during the first half of the 20th century, individuals or a small number of researchers performed most ecology studies over a short time period and with limited funding. Other than the data published as tables in a manuscript, data sharing was not the norm. Few, if any, data collection and data management standards existed or were followed for documenting (i.e., ascribing metadata), quality assuring (i.e., quality assurance/quality control; QA/QC), and organizing (i.e., database management) data. This situation began to change in the 1960s in response to the emergence of "big ecology" (sensu Coleman, 2010) programs that followed in the footsteps of the International Geophysical Year of 1957–58, an international earth sciences research effort that included a focus on meteorology and oceanography.

Coleman (2010) provides a detailed history of many of the large ecological and environmental research programs from the 1950s through the present including the International Biological Program (IBP), the Long-Term Ecological Research Program (LTER) and International LTER Program (I-LTER), and the National Center for Ecological Analysis and Synthesis (NCEAS). The timeline and characteristics of these and other programs that extend to the present day (i.e., Global Biodiversity Information Facility (GBIF), National Ecological Observatory Network (NEON), and Ocean Observatories Initiative (OOI)) are presented in Table 1. The included programs are similar in that the U.S. National Science Foundation partially or wholly funded them and they reflect the transition from short-term (i.e., 1–3 years), low-cost, minimally-staffed projects to long-term (i.e., decade or longer), high-cost, multi-institutional and multi-national projects that serve a large group of