DATA MANAGEMENT FOR LTER:

1980 – 2010

A POSITION PAPER

prepared by

ROBERT J. ROBBINS

in conjunction with the NSF thirty-year review of LTER

2011

DATA MANAGEMENT FOR LTER: 1980 – 2010

CONTENTS

DATA MANAGEMENT FOR LTER: 1980 – 2010

INTRODUCTION

The Invisible Present

All of us can sense change — the reddening sky with dawn's new light, the rising strength of lake waves during a thunderstorm, and the changing seasons of plant flowering as temperature and rain affect our landscapes. Some of us see longer-term events and remember that there was less snow last winter or the fishing was better a couple of years ago. But it is the unusual person who senses with any precision changes occurring over decades. At this time scale, we are inclined to think the world is static, and we typically underestimate the degree of change that does occur. Because we are unable directly to sense slow changes and because we are even more limited in our abilities to interpret their cause-and-in effect relations, processes acting over decades are hidden and reside in what I call the invisible present.

Magnuson, John J. "Long term Ecological Research and the Invisible Present." *Bioscience* 40, (1990): 495-501.

Revealing the invisible present is the goal of the Long-Term Ecological Research (LTER) program. Hundreds of scientists, working at more than two-dozen sites study phenomena that unfold over decades (or longer). As this requires long-term continuity of effort, funding for participating LTER sites comes through special mechanisms that allow for activities to continue uninterrupted far beyond the typical three-year award. To receive such funding, LTER sites and staff must agree to participate in a larger LTER Network, with a commitment not only to producing excellent local science, but also to providing support (largely in the form of shared data sets) for others to carry on the work, at other places and in other times.

In addition to the regular review of individual LTER sites, every ten years NSF has convened a review body to consider the LTER program as a whole. This position paper was prepared in conjunction with the thirty-year review.

When the committee started its review, Joann Roskoski, the acting AD of the Directorate for Biological Sciences, urged us to think big:

> Don't just examine LTER at 30, think about LTER at 100! Imagine what could be learned from 100 years of LTER findings, then ask whether LTER at 30 is on track to deliver the goals for LTER at 100.

That vision has guided my thinking about the current state of LTER, especially with regard to data issues. In the section below, my consideration of today's "data issues" will largely be in the context of LTER@100 — that is, LTER sufficiently far in the future that none of the LTER's founding scientists, and few of today's, will still be alive, much less practicing research. Any value that LTER@30 provides for LTER@100 will come in the form of published findings and shared long-term data sets.

The analysis in this position paper is confined to data issues related to secondary use (also known as "third-party" use) — the use of data sets by individuals not associated with their original collection. Because effective primary use of data is directly related to the science being conducted at individual LTER sites and is well assessed as part of the science review during sites' individual competitive renewals, it is not considered here.

DATA ISSUES

Methods for the Review of Data Issues

To assess issues related to LTER support for the secondary-use of data, I visited the LTER Network Office (LNO) and conducted extensive interviews with many LNO staff; visited several individual LTER sites; interviewed several individual LTER site data managers; interviewed several NSF staff; interacted directly with the LTERNET data portal and with individual LTER site web sites to assess the availability, accessibility, and usability of data sets; reviewed the literature on LTER data management; and considered the state of data sharing and availability for the biological sciences in general.

Access to Data

Like any science, long-term ecological research depends upon access to high-quality data. All fields have the need to collect, manage, and share data, but ecological research is especially dependent upon *historical* data. If laboratory data are lost, one can, if necessary, always repeat a laboratory experiment. But, if historical data are lost, they can be lost forever — it is simply impossible to take yesterday's measurement today. Even if historical data are preserved, their meaning can be lost if they are not accompanied with sufficient metadata — data about the data — to make them interpretable. Figure 1 shows what happens over time to unmanaged, undocumented data. Irreversible degradation inevitably occurs until finally there is no residual value left.

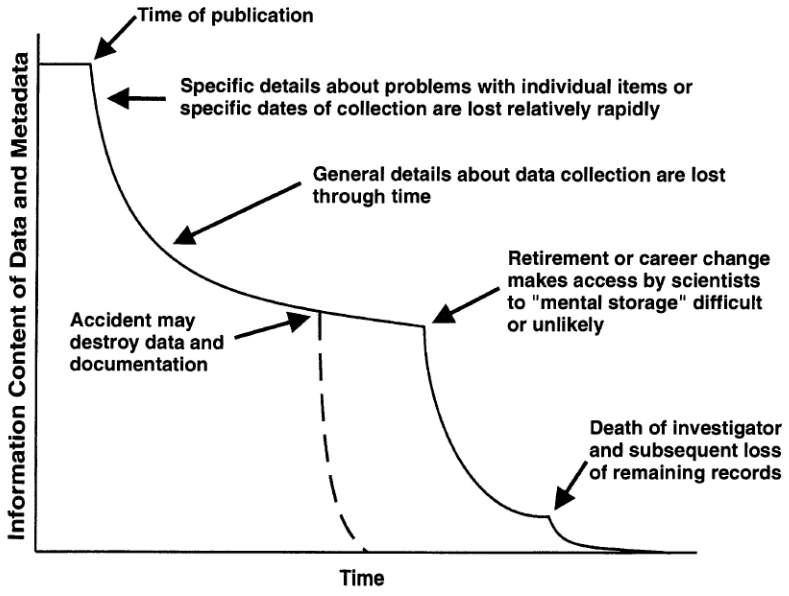


Figure 1 Example of the normal degradation in information content associated with data and metadata over time ("information entropy"). Accidents or changes in storage technology (dashed line) may abruptly eliminate access to remaining raw data and metadata at any time. (Figure taken from Michener *et al.*, 1997)

Information entropy can only be resisted through active intervention. Initially the data must be checked for quality, then appropriately annotated for future use. The data must be stored on media resistant to degradation, and the state of the data must be checked

periodically. Regular data migration to usable media (and formats) is required. QA/QC activities must occur periodically. And, it is essential that the data be *used*, at least occasionally. Without actual use, it is impossible to assess the data's true *fitness for use*.

Michener *et al*. (p. 333) note that all of this is neither easy, nor cheap:

> Real costs are associated with editing data and metadata and making them available to the scientific community in hard-copy or electronic formats. Research grants and other existing funding mechanisms are often insufficient to support development of a comprehensive set of metadata.

Furthermore, if the goal is to make the data available essentially in perpetuity, who should bear the long-term responsibility for hosting, managing, and providing the data *after the funded project is completed*? Clearly it cannot be the original researcher, if the LTER@100 goal is to make data available not only beyond the end of the funded project, but also beyond the life-span of individual scientists or even of individual institutions.

From the perspective of LTER@100, truly long-term responsibility for data management must reside with some third-party, for by definition it cannot be done by the original scientist.[1]

It is well known that the complexity (and associated costs) for managing data to meet the needs of *every* user will be far greater than that required by *any one* user. Trying to balance utility versus cost is a constant challenge for data managers, who frequently find themselves dealing with a *Goldilocks Effect*, where some users consider their efforts inadequate, others find them excessive, and only a few judge them to be just right. This is captured in Figure 2, derived from the Michener *et al*. paper.

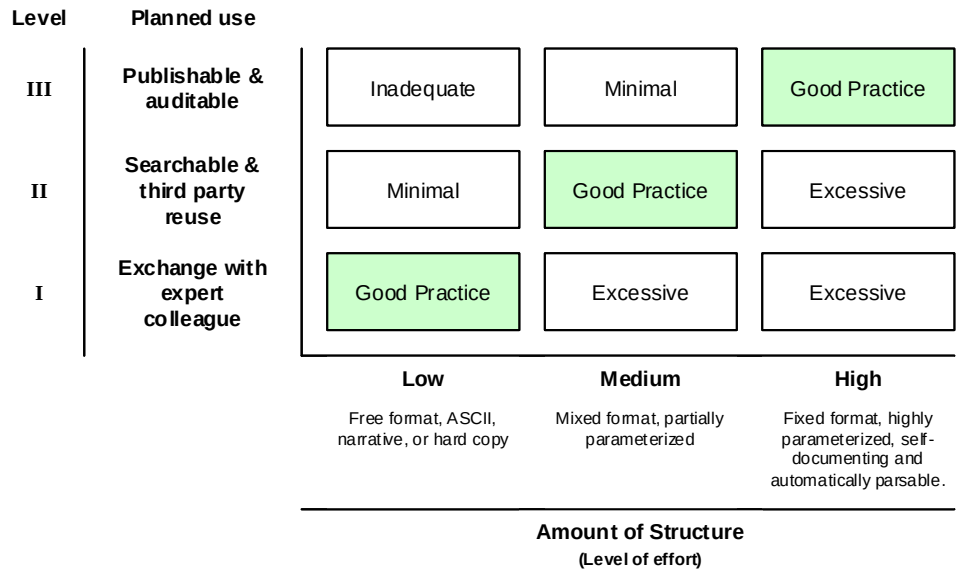


Figure 2 The degree of metadata format and structure necessary for different levels of projected secondary data utilization. (Figure taken from Michener *et al*., 1997)

[1] Nor is it reasonable to expect that original grantee institution would maintain the data, unless that were made a formal condition of the award (which might require that LTER awards be made as contracts, or cooperative agreements). Even then, it is difficult to see how institutions could reasonably be expected to incur unfunded costs forever, just because they once received an LTER award.

Note that the Goldilocks Effect will be most pronounced for systems exhibiting the *medium* level of structure. Those at the low end will hear of satisfaction from Level I users, but complaints of inadequacy from all others. Those at the high end will hear of satisfaction from Level III users, but may experience complaints of technical overkill from the rest. If, however, the high end is deployed with sufficient attention to ease of use and simplicity in the interface, tremendous complexity may be hidden beneath the surface and yet elicit few complaints of overkill. This might be called the iPod Effect.

At the very beginning of LTER, the primary goal was to exchange data with expert colleagues and the required data-management effort was correspondingly low. Today, LTER is operating in the middle zone, where the data are searchable and available for third-party use (but are not yet quite ready for seriously demanding third-party use). To achieve the vision for LTER@100 and support Level III users, LTER data-management practices will have to move to the far right of the diagram. Already LTER data management is experiencing the Goldilocks Effect. Each of the following statements reflects a sentiment that I have encountered during the preparation of this paper:

I. I don't understand why the data managers always want to build some big, complicated system, when all that is really needed is that the data be available. It's like they just want to build computer stuff, without regard for the actual science.

II. Many databases are online; however, choosing the best ones can be a difficult task. We found the LTER sites to be the best for downloading and interpreting data. These sites contain a plethora of data from many different types of ecosystems. LTER's various sites share a common goal of promoting ecological science by fostering the synergy of information systems and scientific research.

III. Aggregating LTER data for a real, large-scale comparative study is difficult and tedious. There are too many manual steps and even then the data may turn out not to be truly comparable for a variety of reasons. It's like they think that just putting files on a web server is enough, without regard for the scientific and computational effort necessary to actually use the files.

Note that these attitudes reflect exactly what would be expected when a Level I, Level II, or Level III user encountered present, medium-effort LTER data systems.

While readily supporting Level III secondary use may be a goal for LTER@100, it is important to note that providing true Level III support is complicated, difficult, and expensive. Many authors have held up DNA-sequence databases as exemplars towards which ecoinformatics should strive. Before attempting to commit ecology and LTER to this path, we should recognize the massive effort that has gone into making DNA-sequence databases so effective. Sequence data are served world-wide from three primary sources, GenBank (in the US), EMBL (in Europe), and DDBJ (in Japan). Combined, the current and past budgets for these three resources far exceeds $500,000,000. These numbers include only the costs to collect, manage, and serve the data. They do not include any of the costs associated with producing the data. All three are (at least in part) government-funded entities, with large permanent staffs.

These budgets can be justified because access to aggregated sequence data fuels not only current basic research, but also practical work in agriculture, energy, engineering, and

medicine. The demand for sequence-data services is high, and their on-line resources provide downloads to millions of individual users daily. New data flow into them at great and ever-increasing speed — every ten days more data are added to GenBank than were added in the first ten years.[2]

From the perspective of LTER@100, it will be important to answer the question, how much will be enough? What functionality will be sufficient to accomplish reasonable scientific goals, while cost-effective enough to be practicable?

With IT technical cost-effectiveness still improving according to Moore's Law, we can expect that substantially more sophisticated systems than are currently available will be practically implementable for LTER@100. Even if current technology is inadequate, future innovations may provide a path forward.

A 2010 paper by Peters (included as an appendix) offers a vision of maximally useful ecological data that begins with original observations, then flows through a series of processing steps that transform the source data to integrated data and then into derived data products, that ultimately support interpretations. A key to achieving that vision will be moving beyond integrated data and on to *derived data products*. Another key is recognizing that useful data can come from a variety of sources, not just LTER, and that the aggregation and derivation steps are also not necessarily restricted to LTER. Figure 3 (taken from Peters, 2010) illustrates the steps in this future process. Note that this presentation outlines an ideal state for *what* should be done, without offering suggestions regarding *who* should do it, or *where* it should be housed. A critical open question is, who should provide the infrastructure necessary to house, manage, and serve the repositories of integrated data and derived data products.

This is a challenge for all science, not just LTER. Repositories for persistent nucleotide sequence data exist in GenBank, DDBJ, and EMBL, but these are very large, very expensive government entities that cannot readily be replicated for other fields. Popular culture offers the model of *crowd-source-managed* repositories, such as YouTube, but these can also be stunningly expensive to own and operate. In 2006, Google purchased YouTube for $1.65 billion and current estimates (Google does not provide details) of YouTube annual operating costs are in the vicinity of $500,000,000.

In short, determining how to host and manage the world's accumulating scientific data is currently a great, and as yet, unsolved problem.

---

[2] In addition to practical budget concerns, there are other reasons why the sequence databases are in fact poor models for other fields to emulate. From a Kuhnian perspective, accumulating DNA sequences is a perfect example of *normal* science, or what has been called *puzzle-solving* or *fill-in-the-blanks* science. With the Watson Crick model, we learned the physical-chemical structure of DNA and now we are just collecting millions of instances of sequences. The data of interest occur as just one type and there are well-established algorithms for manipulating the data. Other fields are collecting data that fall more into the Kuhnian pre-paradigm model where the boundaries between known and unknown are fuzzier and where the need for complicated metadata is great. To understand the data management needs of LTER@100, imagine what a database of information on the chemical nature of the gene would have been like in 1950, before Watson and Crick provided the actual structure. Some believed that the gene must be made of proteins, some considered that it might consist of DNA, while others argued that DNA had a simple-minded repetitive tetranucleotide structure that could not possibly encode information. In this confusion, some attempted descriptions of the necessary chemical structure of the gene — aperiodic crystals — invoked terms that were more metaphysical than scientific. Building a large, scalable database to support the needs of Kuhnian normal science is hard. Building a database to support the needs of discovery science — e.g., long-term ecological research — is much harder.

Box 1. A framework for science-driven synthesis

A developing framework for synthesis includes five major steps and four classes of products that will result in making complicated data easily accessible for understanding and prediction by a broad audience (Figure I).

First, three strategies of ecological research (long-term studies; short-term, process-based studies; and broad-scale observations) result in large amounts of source data collected by individuals, sites, and networks of sites in a variety of formats, units, temporal and spatial resolutions, and degrees of complexity that need to be assembled. These data are often variable in their quality in terms of the degree to which they have been checked and corrected for errors.

Second, these diverse datasets need to undergo quality assurance and control, and to be standardized and integrated into one database, either a virtual database with internet links or a physical database. Much will be learned about the structure of diverse datasets that will provide important feedbacks to the data collection process.

Third, these data need to be converted into common aggregations to simplify their temporal and spatial resolutions that will allow comparison across sites and studies, and to promote synthesis. Derived data products need to be created, including X–Y graphs, maps, animations, and statistical results. The aggregation and analysis process will require new software tools and quantitative analyses, and training of scientists and information specialists to use and develop these tools.

Fourth, these derived data products need to be combined with other knowledge sources, new technologies, and approaches to promote new interpretations and synthesis of the data. A broad user community will be needed that includes individuals (e.g. scientists, land managers, citizen scientists, and information managers), networks of sites (e.g. LTER, USDA, and NEON), synthesis centers (e.g. National Center for Ecological Analysis and Synthesis [NCEAS, http://www.nceas.ucsb.edu/], National Evolutionary Synthesis Center [NESCent, http://www.nescent.org/], National Institute for Mathematical and Biological Synthesis [NIMBioS, http://www.nimbios.org/], and Powell Center; http://powellcenter.usgs.gov/), and state and federal agencies working together. These activities need to provide important feedbacks to the collection of additional data as well as to the development of tools and expertise for future analyses.

Fifth, these interpretations will need to inform policies, practices, and actions, and provide feedbacks to the collection of additional data. New technologies will need to be developed, and training of scientists and information managers in synthetic research will be needed to meet the challenges associated with synthesis.

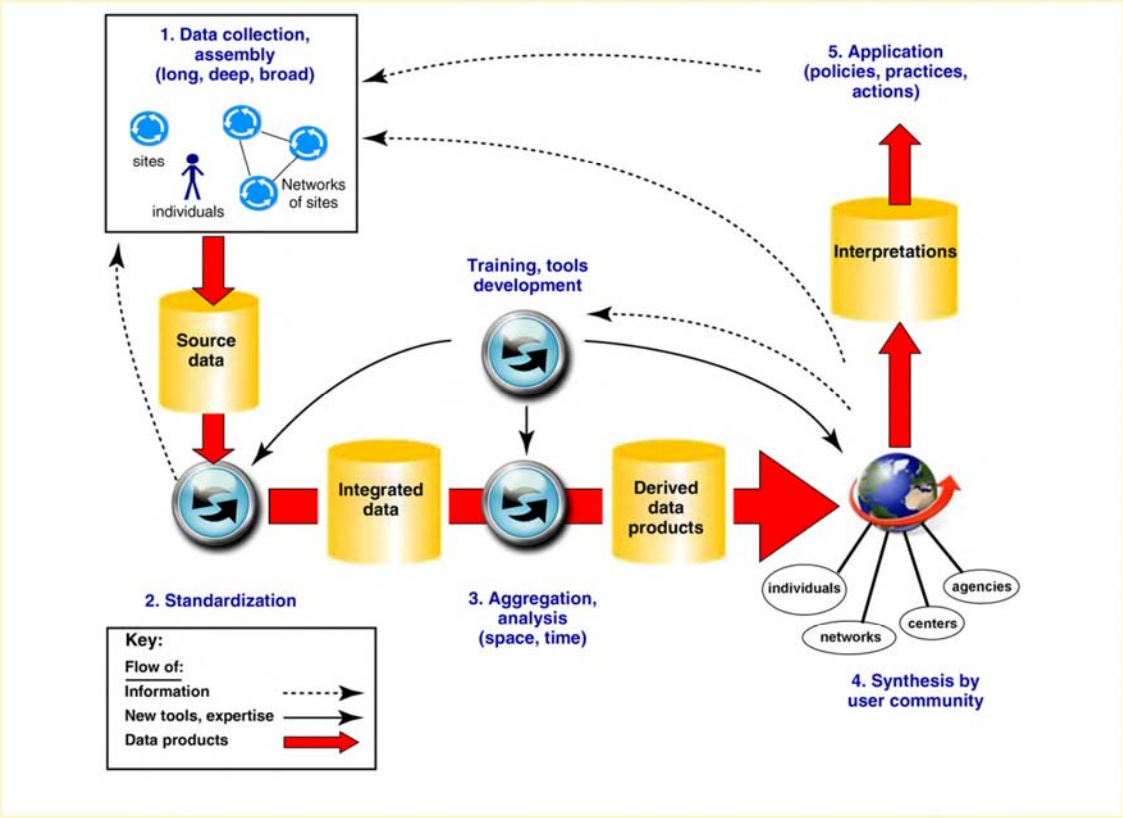


Figure I. A developing framework for synthesis includes five major steps (blue text) and four classes of products (yellow cylinders).

Figure 3. Peters' vision (Peters, 2010) for a future flow of alternating, value-adding processing steps and repositories to transform raw observations into integrated data and then into derived data products, ultimately yielding information useful for secondary-use ecological research, policy making, and other valuable activities.

Data Sharing

Data Sharing is Not a Natural State

Data Sharing — Not a New Problem

[Isaac] Newton ... clashed with the Astronomer Royal, John Flamsteed, who had earlier provided Newton with much needed data for *Principia*, but was now withholding information that Newton wanted. Newton would not take no for an answer; he had himself appointed to the governing body of the Royal Observatory and then tried to force immediate publication of the data. Eventually he arranged for Flamsteed's work to be seized and prepared for publication by Flamsteed's mortal enemy, Edmond Halley. But Flamsteed took the case to court and, in the nick if time, won a court order preventing distribution of the stolen work. Newton was incensed and sought his revenge by systematically deleting all references to Flamsteed in later editions of *Principia*.

Hawking, Stephen W. 1988. *A Brief History of Time*. New York: Bantam Books. p. 181

There are social as well as technical problems associated with data sharing. In other scientific communities (*e.g.*, genome research), establishing a culture of data sharing required active social-engineering efforts on the part of database staff, publishers, funding agencies, and community leaders. At first, researchers did not actively share their data — GenBank staff manually copied the data out of published papers. As sequencing increased, GenBank began to fall significantly behind the published literature. The solution was a combination of technical and social engineering, involving the direct submission of data from the researcher to the database, in a format that could be automatically accessioned by the database. This required a technical solution (standard data formats had to be developed and a tool made available to the researcher to facilitate the creation of a direct submission) and a social solution (the researchers had to be motivated to take the effort to submit the data).

IntelliGenetics, Inc., (the company then holding the prime contract for GenBank) and staff at the Los Alamos National Laboratory (the site holding the content-creation portion of the GenBank contract) developed *AuthorIn*, a software package that allowed researchers to submit data directly to GenBank. This solved the technical problem. As sequencing technology improved and researchers produced longer sequences, the idea of printing them verbatim in journal articles became untenable. Many journals adopted the policy that they would (a) no longer print sequences in the journal and (b) only accept a sequence-related paper for publication if the authors could demonstrate that the relevant sequence data had already been submitted to GenBank. This addressed the motivation issue.

Choosing not to print sequences was a natural, self-interested choice for publishers wishing to reduce printing costs, but many saw the submission-before-publication requirement as problematic: "If I implement this policy, but Journal X does not, some of the best scientists may choose only to publish in Journal X, because they do not like having to comply with the data-submission policy." Several journals did not spontaneously adopt the data-submission requirement, and some journal editors had to be lobbied vigorously and for some time before finally agreeing to a formal policy.

Other fields have seen similar problems overcome through active social engineering. The simple fact is, data sharing is not a natural condition. Much effort is put into producing a data set and most researchers are loathe to part with *their* data. After all, data are the raw material out of which scientific discovery, and thus scientific careers, are made. A frequently expressed concern is, "I don't want to share my data and then have someone else make a major discovery out of my data."[3]

The problem is exacerbated if researchers are expected to devote significant additional effort to making their data available to others. As Michener *et al.* (1997) have noted, "Although increasing metadata structure (*i.e.*, format definition) reduces the burden on data re-users, it significantly increases the burden on the data originator."

The problem is so common in modern science that the journal *Nature* had data sharing as the theme its 10 September 2009 issue. This illustrates that there is nothing special about data-sharing problems in long-term ecological research. The problem is pervasive and can only be resolved through active intervention, especially by funding agencies and publishers. An editorial in that special issue — Data's Shameful Neglect — is important and highly relevant to LTER data issues and is reprinted on page 9.

Scientific Careers are based on Discovery, not Service

The Measure of Scientific Success

Fail to discover, and you are little or nothing in the culture of science, no matter how much you learn and write about science [nor how much you do in the service[4] of science]. Scholars in the humanities also make discoveries, of course, but their most original and valuable scholarship is usually the interpretation and explanation of already existing knowledge. When a scientist begins to sort out knowledge in order to sift for meaning, and especially when he carries that knowledge outside the circle of discoverers, he is classified as a scholar in the humanities. Without scientific discoveries of his own, he may be a veritable archangel among intellectuals, his broad wings spread about science, and still not be in the circle. The true and final test of a scientific career is how well the following declarative sentence can be completed: *He (or she) discovered that…*

Edward O. Wilson. 1998. *Consilience*. New York: Alfred A. Knopf. p 56

Although data sharing may have become of critical importance for science, it is still not a natural component of the scientific enterprise. This is not because of any inherent selfishness on the part of scientists, but because successful scientific careers are measured by discovery, and little else.

[3] This concern is especially strong, if the data were collected as part of a research project, in which success would be measured by discoveries made, not data shared. When the human genome project went into final-stretch mode, many large-scale sequencing centers were funded with the express goal of producing bulk sequence that would be shared immediately and interpreted by others. This kind of purely data-driven (as opposed to hypothesis-driven) effort is rare in science.

[4] Several years ago, at a meeting of the International Advisory Committee to the International Nucleotide Sequence Databases (the committee of scientists that collectively advise GenBank, DDBJ, and EMBL) a member of the database staff (who happened to hold a PhD in biology), started to make a point by saying "As a scientist myself, I…" when he was interrupted by a European member of the committee, who announced in a booming voice, "You are not a scientist, you are a servant of science!"

Data's Shameful Neglect

More and more often these days, a research project's success is measured not just by the publications it produces, but also by the data it makes available to the wider community. Pioneering archives such as GenBank have demonstrated just how powerful such legacy data sets can be for generating new discoveries — especially when data are combined from many laboratories and analysed in ways that the original researchers could not have anticipated.

All but a handful of disciplines still lack the technical, institutional and cultural frameworks required to support such open data access (see pages 168 and 171) — leading to a scandalous shortfall in the sharing of data by researchers (see page 160). This deficiency urgently needs to be addressed by funders, universities and the researchers themselves.

Research funding agencies need to recognize that preservation of and access to digital data are central to their mission, and need to be supported accordingly. Organizations in the United Kingdom, for instance, have made a good start. The Joint Information Systems Committee, established by the seven UK research councils in 1993, has made data-sharing a priority, and has helped to establish a Digital Curation Centre, headquartered at the University of Edinburgh, to be a national focus for research and development into data issues. Other European agencies have also pursued initiatives.

The United States, by contrast, is playing catch-up. Since 2005, a 29-member Interagency Working Group on Digital Data has been trying to get US funding agencies to develop plans for how they will support data archiving — and just as importantly, to develop policies on what data should and should not be preserved, and what exceptions should be made for reasons such as patient privacy. Some agencies have taken the lead in doing so; many more are hanging back. They should all being moving forwards vigorously.

What is more, funding agencies and researchers alike must ensure that they support not only the hardware needed to store the data, but also the software that will help investigators to do this. One important facet is metadata management software: tools that streamline the tedious process of annotating data with a description of what the bits mean, which instrument collected them, which algorithms have been used to process them and so on — information that is essential if other scientists are to reuse the data effectively.

Also necessary, especially in an era when data can be mixed and combined in unanticipated ways, is software that can keep track of which pieces of data came from whom. Such systems are essential if tenure and promotion committees are ever to give credit — as they should — to candidates' track-record of data contribution.

Who should host these data? Agencies and the research community together need to create the digital equivalent of libraries: institutions that can take responsibility for preserving digital data and making them accessible over the long term. The university research libraries themselves are obvious candidates to assume this role. But whoever takes it on, data preservation will require robust, long-term funding. One potentially helpful initiative is the US National Science Foundation's DataNet programme, in which researchers are exploring financial mechanisms such as subscription services and membership fees.

Finally, universities and individual disciplines need to undertake a vigorous programme of education and outreach about data. Consider, for example, that most university science students get a reasonably good grounding in statistics. But their studies rarely include anything about information management — a discipline that encompasses the entire life cycle of data, from how they are acquired and stored to how they are organized, retrieved and maintained over time. That needs to change: data management should be woven into every course in science, as one of the foundations of knowledge.

Figure 4 Data's Shameful Neglect, an editorial reprinted from the 10 September 2009 issue of *Nature.*

Wilson's observations about discovery being the only measure of a scientific career are especially relevant to data-sharing. To the extent that data-sharing is seen as a *service* function, it is not directly related to personal discovery and thus does not contribute to enhancing a scientist's career. This is a simple fact that cannot be wished away. To get scientists to become more actively involved in the creation and sharing of data sets for secondary use, appropriate incentives for "good" behavior must be put in place (perhaps augmented by appropriate disincentives for bad behavior). This is best done by creating an environment where data-sharing actions (or inactions) have clear consequences.[5] Given that the desired behavior — active engagement in data sharing — is not a natural part of the scientific process, the incentives must come from outside the immediate scientific community, with funding agencies the best placed source of such incentives.[6] This is why it is so important for NSF to increase the importance of informatics in its assessment and review of LTER activities (cf. the discussion below of recommendations from the 20-year review).

Data Sharing Challenges for Long-Term Ecological Research

Metadata are Essential

The most important reason to invest time and energy in developing a metadata is that human memory is short. If data are to undergo any secondary usage, then adequate metadata will be required even if that secondary usage consists of reuse by the data originator.

Michener et al., 1997. Nongeospatial Metadata for the Ecological Sciences. *Ecological Applications*, 7: 330-342.

It is not enough to simply make raw data available for others to use. The data must be sufficiently documented, *i.e.*, accompanied with metadata, to make them interpretable and valuable by someone with no first-hand knowledge of how or why the data were collected. This requires substantial effort on the part of scientists, not just on the part of IT staff. To motivate such behavior, funding agencies must put in place appropriate incentives (and disincentives). They must also make it possible for researchers to share data, by creating tools to facilitate data sharing and by supporting repositories where shared data may be housed.

In the sequence community, this was accomplished by making available data-submission tools (*e.g.*, *AuthorIn*, later *Sequin* and *Bankit*) and data repositories (GenBank). Similar tools and services will be required, if widespread data sharing is to occur in the long-term ecological research community. The adequacy of current LTER activities will be explored below.

---

[5] As any parent knows, no amount of moral exhortation is as effective in getting one child to share fairly the last piece of cake as is the simple rule, "one cuts, the other chooses." When the first child grasps the consequences of his action (an unequal slicing of the cake will result in his getting the smaller piece), he divides the piece with microtome-like precision. In management speak, this involves *structuring the reward system to ensure the alignment of incentives*.

[6] Journals are another potential source of leverage. The two primary rewards in a scientific career are (1) receiving research funding and (2) being published in a reputable journal. The fact that these rewards come from outside the institution in which researchers are employed creates an odd dynamic, not unlike the issues associated with the third-party-payer dynamic in American medicine. This is too wide-ranging and two complicated a subject to explore here, except to note that funding agencies and scientific publishers play huge roles in rewarding the behavior, and thus in shaping the behavior of scientists. In general, the policies of funding agencies and publishers have a substantially greater effect on scientists' behavior than do the policies of their home institutions.

RECOMMENDATIONS FROM 20-YEAR REVIEW

The full text for all of the recommendations from the 20-year review of LTER are provided in Appendix I. Several of those recommendations were, either in whole or in part, relevant to achieving the data goals for LTER. The data-relevant components of the recommendations are presented below, along with comments on the extent to which these recommendations have been achieved in the intervening ten years. The recommendations are grouped by target, that is, by whether they were primarily aimed at NSF, at LTER, or at both together.

NSF Should:

(9) *revise proposal guidelines and review criteria to provide greater balance between site-specific and cross-site activities.*[7]

In the 20-year review document, it was asserted that accountabilities and their approximate weighting for LTER site renewal proposals and site visits were

- site-specific research, 50%;
- site-specific information management, 20%;
- site-specific management and governance, 10%;
- cross-site activities, synthesis, outreach, 10%; and
- network-level activities, 10%.

and on that basis, the 20-year review called for greater balance between site-specific and cross-site activities.

Assuming 20% effort on data management and 20% effort on cross-site and network activities, there is 4% effort on cross-site data-management activities. Further assuming an average of 1.5 data-management FTEs at a typical LTER site, there is approximately 0.06 FTEs dedicated to cross-site data-management activities. That's less than a half hour per day, raising the reasonable question, "Just how much cross-site data-management work can get done, if 26 people every day devote half their lunch hour to the project?" The answer, obviously, is not much. The fact that a great deal of cross-site data-management has historically gotten done, despite these numbers, shows that in the past many site-based informatics staff have devoted far more time to cross-site data-management issues than would be justified based on the review criteria cited in the 20-year review.

Representatives of NSF have stated that these specific percentages are no longer in use, having been replaced with a less quantitative approach.[8] Removing explicit

---

[7] A discussion of this recommendation is included in the data-issues section because support for secondary-use of data is a type of cross-site activity.

[8] The 2011 renewal proposal guidelines provided by NSF to LTER sites reads:

The primary review criterion will be scientific merit (NSF's Criterion 1). **Information management** and technology, **site management** (including personnel, fiscal, administrative, institutional and logistical issues), **network participation** (including cross-site, non-LTER, or international research, and involvement in other network activities) and **outreach/education** (training of undergraduate and graduate students, K-12 Schoolyard, application of results to policy and management, etc.) are also important aspects of all LTER projects that will be addressed during the review. Each of these five criteria are evaluated with respect to quality, productivity and

quantification, however, does not necessarily change the established perceptions of grantees and reviewers.

Only one competition for new LTER sites (for coastal ecosystems) has occurred since the recommendations of the 20-year review committee. In that solicitation (NSF 03-599), there was some emphasis on cross-site activities, but certainly not on a par with the focus on local site excellence. One sentence in the narrative instructions of the solicitation reads, "Outline any regionalization, cross-site, or other collaborative efforts involving the LTER network that are planned if they are not part of your core program (network activities)." A reasonable interpretation of the word *any* in this context is that such activities are optional, not mandatory. A more injunctive sentence, would read, "If they have not been included as part of your core program (network activities), describe your regionalization, cross-site, or other collaborative efforts involving the LTER network here."

Conclusion: if NSF desires more cross-site efforts from the LTER network, then more explicit guidance from NSF is still needed. And, the guidance must have teeth — *e.g.*, inadequate network participation will result in loss of funding, regardless of how good the local science.

(11) *increase the importance of informatics in its assessment and review of LTER activities*

Using the coastal-ecosystems solicitation and the 2011 renewal guidelines as references, we see that proposals may contain up to 35 pages (narrative plus figures) describing the proposed science and up to 4 pages (narrative plus figures) describing the "data and information management system and metadata standards to be used". If space allocations reflect importance, then this 10-15% allocation matches the historical emphasis on data management that has been in place since the beginning of the program.

Conclusion: if NSF desires more data-management and data-sharing efforts from the LTER network, then more explicit guidance (with teeth) from NSF is still needed.

(11) *support the informatics core function at a level sufficient to achieve the LTER program's informatics objectives in an aggressive timeframe.*

On the one hand, this is simply a call for more funding and, as such, is indistinguishable from such a plea from any scientific field. On the other hand, it is a call to match program spending with program goals. That is, if informatics is to become of much greater significance in LTER, and if this is to happen on an aggressive time frame, then aggressive allocation of appropriate resources will be required. If aggressive allocation of resources cannot be done, then the goal of significant change on an aggressive timeline must be appropriately modified.

The last ten years have been difficult ones for federal funding of research, with success rates in many programs falling to historical lows. As a result, it is not

---

impact. Research both within and across sites essentially comprises Criterion 1. The "education and outreach" portion is essentially Criterion 2. IM and site management (including non-research aspects of network participation) are part of the review criteria for all NSF Centers, including LTER.

surprising that NSF was unable to respond immediately with aggressive spending for LTER informatics. However, NSF did take advantage of the recent, one-time ARRA funding and has used that to provide a substantial informatics supplement to the LTER network office. Thus, this recommendation was ultimately followed, albeit on a better-late-than-never schedule. As will be seen below, however, some additional funding would still be very helpful in achieving even more significant, network-wide improvements.

(27) *(a) allocate two program officers to administer LTER, with a permanent one in charge of the program and (b) establish a formal, cross-directorate committee of program officers should be established to coordinate LTER funding and program management.*

Although NSF has established a cross-directorate committee to coordinate LTER activities, the goal of having a permanent staffer in charge of overall LTER activities has not been achieved. LTER has seen several recent changes in program staff. This is problematic for a program that is expected to span decades, and it is especially problematic for data sharing, where the goals transcend individual awards and may take many years of coordinated effort to deliver success.

Conclusion: To ensure stability and continuity in LTER operations, especially in the area of data management, NSF should take steps to minimize the turnover of program staff assigned to LTER.

LTER Should:

(2) *become information-driven.*

The full context for this recommendation is that, "LTER science be multidisciplinary, multi-dimensional, scalable, information driven, predictive and model based, education oriented, and increasingly virtual and global." Progress in LTER science over the last decade has followed these recommendations. The field is substantially more interdisciplinary and technology has made it more scalable and information driven.

(10) *establish informatics as a core function by implementing a systemic information infrastructure.*

Since the recommendations were made, LTER has adopted EML[9] (ecological metadata language) as a standard for documenting LTER data sets. Until recently, resources were not available to pursue the development of a true "systemic information infrastructure". However, the current ARRA funding to the network office has allowed a significant effort to get underway to develop and deploy an LTER information infrastructure, based on a Provenance Aware Synthesis Tracking Architecture (PASTA). PASTA is designed to dynamically harvest and archive site-based data and metadata and to use that harvested information to produce synthetically derived data products. These derived data products will then be available through multiple user and machine interfaces, with all derived data described by an associated rich and structured EML document (which will

[9] http://knb.ecoinformatics.org/software/eml/

emphasize the product processing history and its origin – hence the notion of being provenance-aware).

Although the development of a systemic information infrastructure was called for ten years ago, funds to pursue this goal just recently became available and the effort to meet the goal is still a work in progress. As will be explained later, there are still some significant unmet opportunities that could usefully be pursued.

(14) *provide a virtual portal to its legacy data*.

In the narrative portion of the 20-year review, it was also noted: "[T]he development and adoption of a single network approach to metadata would allow researchers to search across the entire set of LTER data holdings, rather than site by site. With a common … metadata format, it would be possible for the network to develop a single gateway, allowing users to search across all LTER data holdings by directing a single query to a single site."

This recommendation has been met. The current LTERNET data portal page aggregates EML-conformant metadata from all LTER sites and allows users to "search across all LTER data holdings by directing a single query to a single site." However, the query only immediately returns information about the existence of available data. Links on to the individual LTER site web pages are often required to access the actual data. The new PASTA system, currently being developed with ARRA funding, will go beyond this and actively harvest shared data, making them available through a single site.

Together, NSF and LTER Should:

(20) tie *LTER's goals and objectives to a realistic budget.*

Tying goals and objectives to a realistic budget means that either the budget must be increased to match the goals, or the goals decreased to match the budget (or some combination). In LTER, as in most grant-funded science, the goals are substantially more ambitious than the budget seems capable of supporting.

For purely scientific activities, this mismatch between budgets and expectations is good (or at least not bad), as truly important science is almost always opportunistic, never quite following in detail plans laid out in the past. The budget mismatch then keeps scientists focused on following only the most promising scientific paths. And, funding-agency program officers are, for the most part, aware that they are overseeing a portfolio of research activities and that success for the portfolio occurs if enough (not necessarily all) of the funded projects yield important results.

For infrastructure activities, however, this is a problem. Infrastructure is invisible when it is successful, becoming visible only when it attracts (negative) attention for some shortcoming.[10] Standard advice in formal project management is:

[10] No one ever started a conversation with, "You know, our drinking water sure tastes pure today," but many are quick to note, "This water tastes funny," if they are unhappy with water quality.

> Success is measured by the ratio of deliverables over expectations. All competent project managers manage deliverables up. Experienced project managers know that they must also take care to manage expectations down.

In the infrastructure components of the LTER program (and that is certainly where most data sharing activities reside), it is crucially important that expectations be managed down to reasonable levels while managing deliverables up. It may be nice to have vaultingly high ambitions for the data-sharing component of LTER, but if that is not matched with equivalently high budgets, the results will be frustration and perceived failure, regardless of how much is actually delivered.

In this review of data issues associated with LTER, some significant mismatches between ambition and resources have been detected. Going forward, both NSF and LTER staff (especially in the Network Office) should take care that data-management expectations are in fact in line with data-management budgets.

FINDINGS ON THE STATE OF LTER DATA

Overview

The short version of my findings is

There is a substantial amount of LTER data available,

The data can sometimes be difficult to find and use, and

The current problems are not unexpected, given the size of the challenge, the limits of current technology, and the resources available.

Significant opportunities for improvement exist.

The Purpose and the Challenge of Long-Term Data

The LTER program was created to allow the study of long-term phenomena that could not be studied effectively over the course of a typical three- or five-year funded project. The mechanism of LTER funding provides the stability needed to address the problems of the *invisible present*. “Long-term” is, of course, a relative term, since it could apply to anything spanning decades, centuries, or millennia. The LTER 30-year review committee was encouraged to think about LTER at 100, so I will employ that time frame here.

If the work of LTER today is to contribute to insights on phenomena spanning multiple decades, or even centuries, it will more likely be from archived data than from the published literature. Thus, the creation and sharing of long-term data sets is clearly an essential part, a *sine qua non*, of the LTER program.[11]

---

[11] Does this suggest that there may be some problems with the priorities in the LTER program, where review and assessment gives more weight to current, site-specific science than to long-term data collection? The answer is no, because knowledge, like any other asset, exhibits time value (the present value of an asset must be discounted, if the asset will not be received until some point in the future). For example, anyone should be willing to pay $20 today to receive $20 today (*e.g.*, giving two ten-dollar bills for a twenty), but what should someone be willing to pay today to receive $20 in the future? Calculating the present value of a future asset is a standard approach used by economists and it depends upon an estimate of the appropriate discount rate, which is more or less equivalent to the estimated average rate of return available over the time period. For example, assuming a 5% discount rate, a rational person should be willing to pay $19 to receive $20 next year, but only willing to pay fifteen cents for the

Such long-term data sets will be valuable only if they are:

> *available*: the data must collected and then stored in a way that they can be retrieved for future use,
>
> *locatable*: archived data sets that cannot be found are of the same value as data sets that never existed,
>
> *accessible*: the data set must be accessible after it is located (a data set stored on obsolete media can be little better than lost data),
>
> *understandable*: the data must be sufficiently well documented so that they can be used sensibly; for example, to compare average daily temperatures across multiple data sets one must know how the averages were calculated — as weighted averages across minute-by-minute measurements (as can readily be done with today's instruments) or as the half-way point between the daily maximum and minimum (as was the only possible with max-min thermometers), and
>
> *usable*: to be truly usable, data sets should be automatically parsable, meaning that it should be easy for software to manipulate unambiguously the individual components of the data set.[12]

Historical Context

With the current ubiquity of computing and networked information, it is easy to forget how recently this information-everywhere phenomenon has occurred. When considering the state of LTER data management we must take care to place past LTER data-management activities in their proper historical context.

For example, LTER was first funded in 1980, before WIFI, before the internet, before generic email, even before the first IBM PC. When LTER started, GenBank did not exist. When Amazon sold its first book and when Microsoft first shipped an operating system with built-in support for networking, LTER was already fifteen years old. It was twenty years old when the DOT-COM bubble popped.

Figure 5 illustrates some events in the history of technology in parallel with events in the history of LTER.

---

right to receive $20 in 100 years. Similarly, if knowledge is not to be obtained for 100 years, its present value must be appropriately discounted.

12 Anyone who has made significant efforts to combine and analyze data from multiple sources, even formally operated sources such as public genome-map databases, knows that many hours of manual labor can be required to get the data cleaned up and reformatted for the common analysis. The problems can be simple (converting units) or complex (pulling structured data out of unstructured text), but the end result is often that several person-weeks of effort may be required before the first analytical run can be attempted.

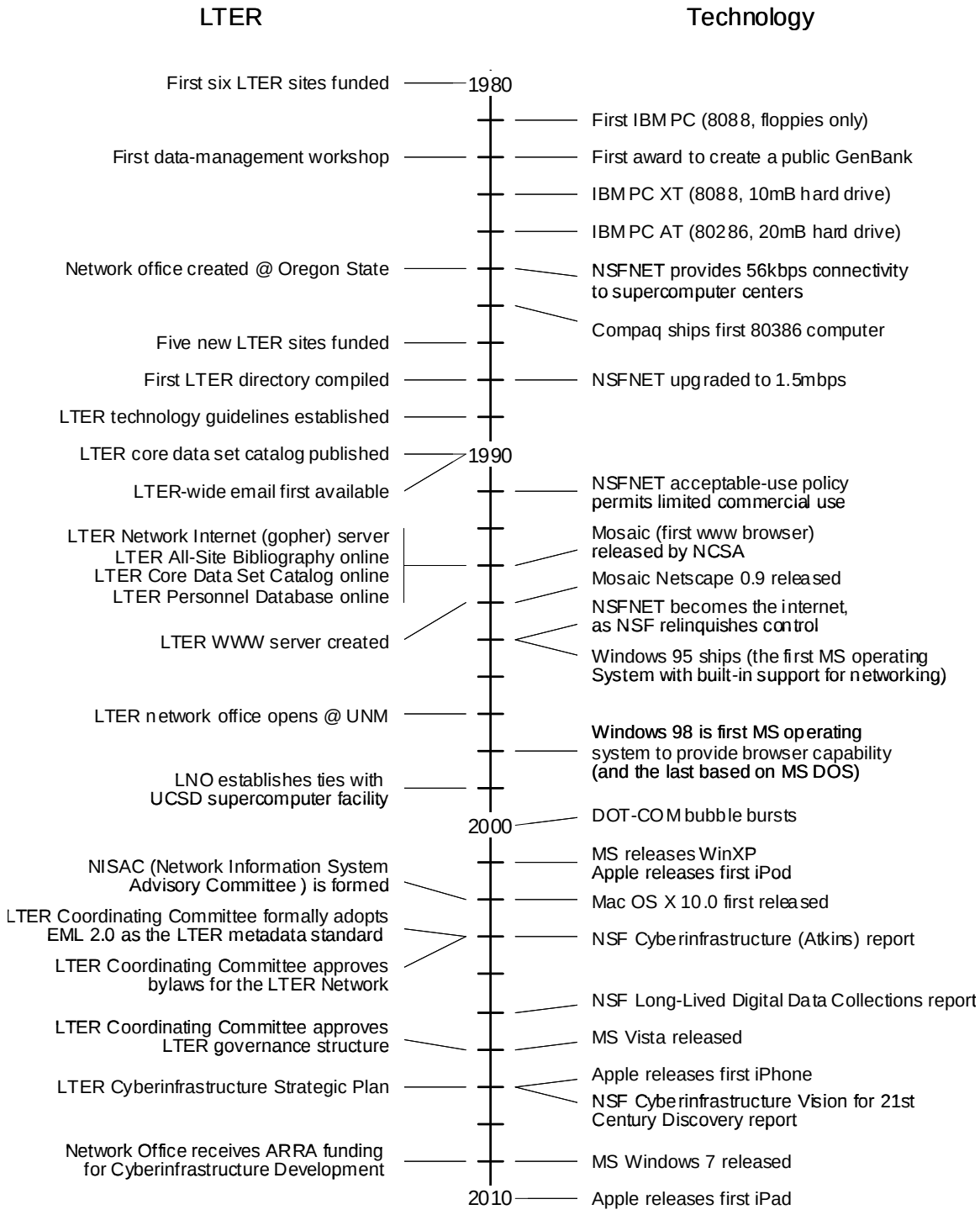


Figure 5. Events in the history of LTER compared with events in the history of technology.

The Challenge

The challenge of successful data sharing in ecology was nicely outlined by Porter and Callahan in 1994:

> The success of ecology as a science depends on development of environmental databases. No single individual is able to collect all the data needed to provide an integrated view of complex ecological systems. Success of shared ecological databases depends on the willingness of investigators and institutions to contribute and use data. However, there is a fundamental dilemma embodied in database creation if benefits derived from shared databases are larger for data users than for data providers. … This inequity, favoring daily users over data contributors, is the fundamental dilemma facing investigator-based environmental databases. It is not reasonable to expect scientists to act against their own best career interests, even if ultimately their actions benefit the scientific community and society in general. … Ecology and evolutionary biology stands virtually alone among the environmental and environment-related sciences in the lack of some agency- or community-mandated data archiving and data-sharing policy. … Long-Term ecological Research (LTER) sites are conspicuous exception to the general lack of information management policies in ecology and evolutionary biology. An early recognition that long-term information management is a critical part of long-term research led to a requirement by the National Science Foundation that each LTER site conduct an active program of data management. NSF left the specific forms of information management programs and policies to the discretion of researchers at individual sites.

Because this *fundamental dilemma* still exists today, data sharing will not occur if merely left to high hopes and best intentions. Instead, solutions need to be designed and implemented, with both technical and social engineering required. As noted earlier, funding agencies and publishers must play a critical and active role in modifying scientists' behavior, because only they have the ability to modify some of the key incentives that shape behavior.

Ecology, with LTER leading the way, has seen substantial improvements since Porter and Callahan summarized the problem. Technology has reduced the burden on the providers and social norms have changed to emphasize the expectation of sharing. In some fields (most particularly DNA sequencing), publication is now sufficiently dependent upon the pre-submission of data to a public archive that the motivation to publish now drives the motivation to share data.[13]

According to Mark Costello (Costello, 2009) there is a crisis in the sharing of environmental data: "Despite policies and calls for scientists to make data available, this is not happening for most environmental- and biodiversity-related data because scientists' concerns about these efforts have not been answered and initiatives to motivate scientists to comply have been inadequate." More recently, several efforts have been made to tie publishing to data submission in ecology (Bruna, 2010; Moore, *et al*., 2010; Rauscher *et al*., 2010; and Whitlock, 2011), although not yet to the extent seen in molecular biology.

Clearly, data sharing is still seen as a problem that is widespread across ecology and evolutionary biology. Over its thirty-year existence, the LTER program has been constantly evolving its approach to address the fundamental dilemma. At its inception

[13] New entries flow into GenBank through direct submission from researchers at a rate in excess of *2500 new entries per hour*, every hour, twenty-four hours a day.

LTER was the only major program in ecological research with data-management and data-distribution policies. Throughout its existence, LTER has been a leader in developing both policies and technologies in support of ecological data sharing.

Data Management in LTER

In the LTER program substantial efforts have been made to make environmental data sets available for others to use. At present, more than 6000 individual data sets are cataloged and locatable via the metacat catalog on the LTER main web site[14] and the LTER Network has adopted a formal data-release policy:[15]

> Data and information derived from publicly funded research in the U.S. LTER Network, totally or partially from LTER funds from NSF, Institutional Cost-Share, or Partner Agency or Institution where a formal memorandum of understanding with LTER has been established, are made available online with as few restrictions as possible, on a nondiscriminatory basis. LTER Network scientists should make every effort to release data in a timely fashion and with attention to accurate and complete metadata.

Porter (2010) provides a history of data sharing in the LTER program. Although Porter's paper is somewhat self-congratulatory in that it emphasizes past success over future challenges, it is also accurate in its assertion that LTER has been a leader in devising both technologies and policies to drive environmental data sharing.

Is the LTER model for data sharing perfect? No. Could it be improved? Yes. But, most importantly, an approach for LTER data sharing is in place and it is generally accepted across the LTER network that data sharing must be the norm.

Although some researchers I interviewed noted problems with accessing and using LTER data, no one asserted that LTER was behind the norm for ecological data and most agreed that no one provides better access to ecological data than LTER. One published summary on the use of databases in the teaching of ecological concepts (LeBare, Klotz, and Witherow, 2000) identified LTER as *the* best online source of ecological data:

> Many databases are online; however, choosing the best ones can be a difficult task. … We found the LTER sites to be the best for downloading and interpreting data. These sites contain a plethora of data from many different types of ecosystems.

According to internally run surveys, in 2005 two LTER sites reported that only half of their data sets were accompanied with metadata, while two others reported even less metadata coverage, and one site admitted that none of its data sets were accompanied with metadata. Now, only one site reports that half or less of its data sets are accompanied with metadata, while twenty-five sites report that 80% or more of their data sets are annotated with metadata.

Given the effort required to generate metadata, this is a significant accomplishment, especially given the relatively low staffing dedicated to information management tasks at a typical LTER site. Current findings from my interviews, plus self reporting from

---

14 http://metacat.lternet.edu/das/lter/index.jsp

15 http://www.lternet.edu/data/netpolicy.html

internal LTER surveys, suggests that the average LTER site only has 1.5 FTEs available to work on data-management issues.[16]

This low-level of IM staffing per LTER site, combined with the large number of LTER sites, produces a double problem: no site has enough staff to meet all of the site-specific and network-wide needs, yet across all twenty-six sites the aggregate of forty or more information-management FTEs constitutes a significant expense.[17]

In the late 1980s, NSF began working with the LTER sites to improve the overall effectiveness of the information managers by encouraging data managers across the network to become acquainted with each other and to see themselves as part of the larger IM community supporting LTER, not just as the lone (or almost lone) data-management person at a particular site. This proved to be remarkably effective when initiated and remains so today. One site information manager told me that attending the annual meeting of the LTER data managers was, from her perspective, the one most valuable event of the year.

Collaborative interactions among data managers from different sites is so effective that recently several spontaneously banded together to pool some of their modest site-specific data-management supplement funds so that they could work together to implement shared tools.

Usage of LTER Data

Although unified statistics are not available, estimates from a sample of sites suggest that more than 10,000 LTER data sets are accessed and downloaded every year, with a substantial portion of the downloads being used by students or for other educational purposes.

Educational users seem to be happy with their data access (*cf.* LeBare et al., 2000), while some research users have expressed frustration with difficulties associated with accessing particular data or with getting data reconfigured for comparative analysis. This is not surprising, given that educators' requirements are less demanding — they are looking for data to *illustrate* findings, not to *produce* findings.

Improving both the usage and the usability of LTER data is an area of significant opportunity.

Interactions with Other Activities

LTER has established working relationships with several other institutions and activities, such as the San Diego Supercomputer Center, NCEAS, DataONE, and NEON. These contacts are good and should be expanded. In particular, LTER should become closely engaged with DataONE as that project matures. As NEON focuses on developing

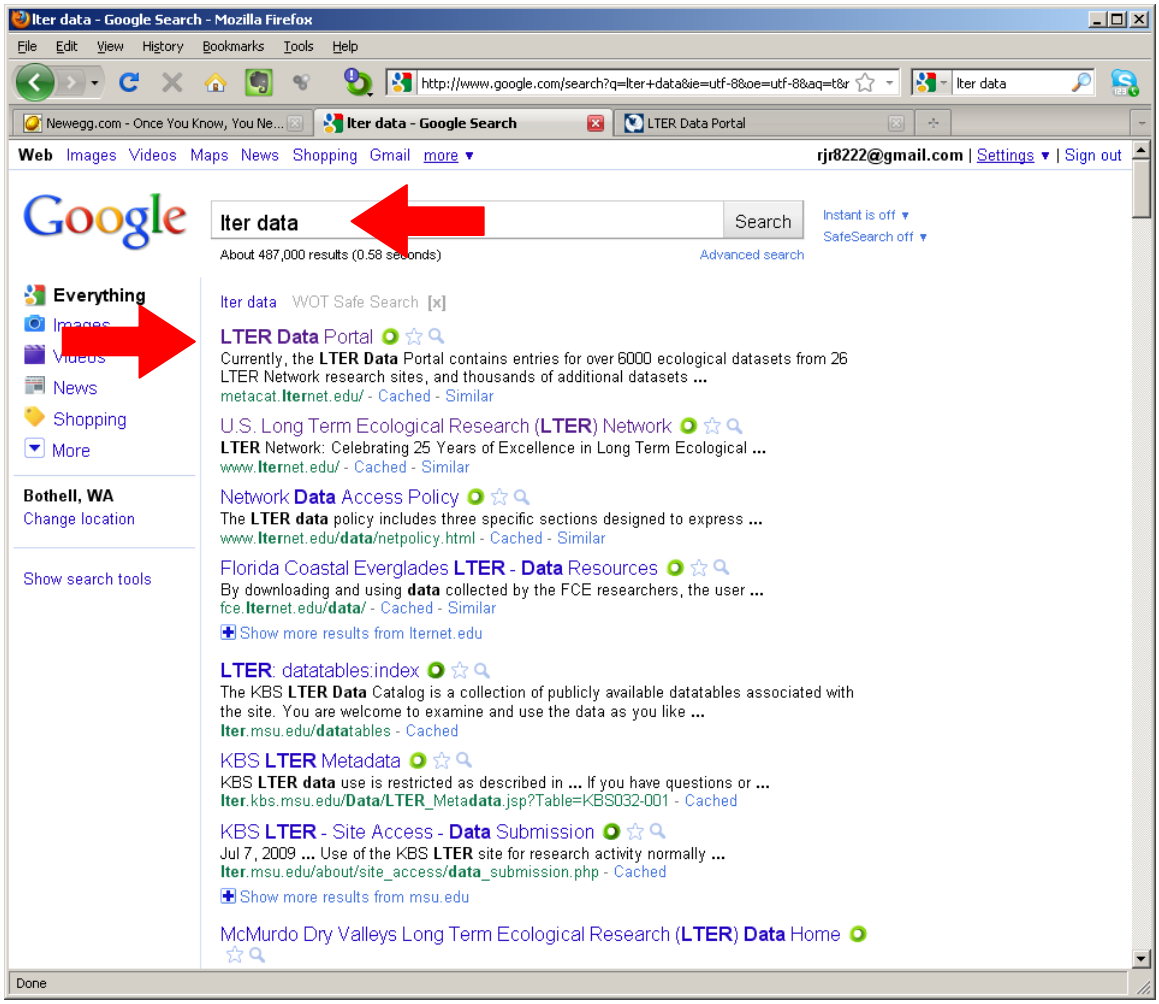
[16] Two sites report atypically high numbers of data-management FTEs. If these two outliers are removed, the average for the remaining 24 sites is 1.36 FTEs.

[17] This problem, and potential solutions, will be discussed in the Analysis and Recommendations section.

advanced infrastructure for continental-scale ecological research, LTER must become engaged with the scientific direction of NEON.[18]

Looking for LTER Data

As part of the review process, I went online to test how easy (or difficult) it is to identify and retrieve LTER data. A Google search on "LTER data" (without quotes) turned up a potentially bewildering 487,000 results, but the very first item on the list was the LTER data portal.[19]



I clicked through to the data portal, then to the "browse" portion of the web site, where I could browse the data sets available from each LTER site. At this point, I started at the beginning (AND: Andrews) and browsed my way through all of the individual sites,

[18] The converse is also true. NEON should become more actively engaged in support LTER. For example, the NEON program is developing aircraft with advanced fly-over remote-sensing capabilities and is developing a schedule of flights that will ensure at least one over-flight for each NEON facility. Given that (a) many LTER sites are also NEON sites and (b) NEON will be acquiring aircraft flight capacity in excess of purely NEON-site needs, serious consideration should be made to including *all* LTER sites in the NEON flight schedules. This merger of LTER science with NEON technology would be highly synergistic and mutually beneficial.

[19] Note that the third item on the list is a link to the LTER data-access policy.

selecting a few data sets at random to see how easy they were to locate, obtain, and understand.[20]

Here I found a catalog of more than 6000 data sets, with an average of approximately 250 data sets per site and a range that was very broad — 1641 described data sets from the Arctic LTER, but only two from Palmer Station. The LTERNET Data Portal provides information from a metadata catalog, based on information provided by individual LTER sites. If a site has many data sets available, but does not provide metadata documentation to the LTER Network Office (LNO), listings for the data will not appear on the Data Portal even though the data could be obtained directly from the LTER site. This seems to be the case with Palmer Station, where a visit to the PAL/LTER website easily found more than 100 data sets.

As I worked my way through the data catalog I encountered variable results. Often I was routed to the web site of the individual LTER site, where the procedures for retrieving the data were locally idiosyncratic. Some sites just made the data available, some required the user to register once and state the intended use of the data, others required a registration and use specification for each data set accessed. Some sites seemed straightforward, others a little trickier to follow. Most ultimately provided data, but a few sites seemed to be suffering from broken links in their cataloged information.

Although it was clear that there is a great deal of ecological data available from the LTER program, it is also clear that the present model of allowing individual sites to be the repository for their own data sets makes for a very tedious experience on the part of users. The fact that the LTERNET metadata catalog is not exhaustive, means that a serious user will spend many hours navigating more than two dozen idiosyncratic web sites just to begin to get an appreciation for what data are available.

Finding LTER Data by Date

I expected to be able to locate LTER data by date fairly easily, given that *Long-Term* Ecological Research is a program with a temporal orientation. Instead, I found that neither the metadata catalog at the LTERNET Data Portal, nor any of the data search pages at any LTER site (except one — North Temperate Lakes) allowed a user to make a direct search for data from a specific date range. I also learned that if a data set has a title that includes a date span, e.g., 1987-1995, that means that the earliest data are from 1987 and the latest data are from 1995, but it does not necessarily mean that the data set includes data from 1990. Given that *time* is perhaps *the* most important variable on long-term studies, this is a problem that needs remediation.

From the perspective of LTER@100, a continued inability to retrieve data sets from specified temporal intervals would be unacceptable.

Despite its apparent conceptual simplicity, structuring large and varied data sets, and their retrieval mechanisms, so that data may be retrieved from, and only from, specified intervals is non-trivial. Nonetheless, this function is critical if LTER data sets are to be of value indefinitely.

---

[20] Not knowing what to expect, I used a software tool — Camtasia — to record the screen and a voice-over narration during the search. A copy of the 75-minute video is available upon request.

North Temperate Lakes site is an outlier in having substantially more data-management staff than the typical LTER site and this is probably not unrelated to the increased functionality and ease of use the user encounters at the NTL web site.[21] This is not surprising, given the simple trade-off relationship described by both Porter and Callalan (1994) and Michener et al. (1997): *The greater the effort by the original data provider, the simpler the task of the secondary data user.*

If the goal is not only to make LTER data sets available, locatable, and accessible, but also understandable and readily usable, then substantially more effort will need to go into data-set preparation than has been possible to date.

Below I present an analysis of my findings and recommendations for future improvements.

ANALYSIS AND RECOMMENDATIONS

Moore's Law and Future Opportunities

LTER is and has been a leader in developing policies and technologies to support the sharing and re-use of ecological data. Despite its past history of success, significant changes need to be made if LTER is to maintain that position of leadership into the 21st century and on to LTER@100. Some of the necessary changes need to occur within LTER, but many need to occur in the relationship of LTER to NSF and to other entities.

In the late 1980s. LTER experienced a major transformation of its information infrastructure. NSF became significantly engaged and (1) helped LTER recognize and maximize the value in its data-management staff, (2) assisted LTER in the acquisition and deployment of GIS technologies, (3) encouraged LTER to specify, and then achieve, a *minimum standard installation* of technology capabilities necessary to be a fully functional LTER site, (4) developed the first all-LTER directory, and (5) emphasized the importance of data management for achieving overall LTER goals.

Another significant transformation in the LTER information infrastructure is needed now — not because of past deficiencies, but because of new opportunities.

Moore's Law continues to drive the evolution of technology and the results of that law — the regular doubling of computer capabilities for constant cost, or the equally regular halving of computer costs for constant performance — gives information technology the ability to transform fields in two markedly different ways. First, when the impossible becomes possible, and second, when the unaffordable becomes affordable.[22]

---

[21] In addition to being able to search by date range, after locating a data set the user is able to sub-select the data set for fields of interest, order the data set by fields of the user's choosing, then retrieve the data in a variety of formats optimized for viewing or computing or parsing. Also, if the user specifies an interest in data from, say, 1930-1950, the NTL interface will retrieve all relevant data sets, even if the span is larger, say 1910-1990, but before the data are extracted and delivered, the user has the option to select just the records from a time period of interest. In the language of relational databases, the NTL system allows the user to extract both PROJECTIONS (i.e., columns) and SELECTIONS (i.e., rows) from the data set of interest, then sort the results. This is very powerful and would greatly reduce the effort required to prepare extracted data for further analyses.

[22] Both factors can be involved — for example, a new technology may be able to transform a field, but the task must first become doable, then affordable given the economics of that field. This is why information-technology-driven transformations have moved through the economy in waves, first affecting activities where cost is no object (e.g.,

The transformation of the 1980s was based on acquiring capability: for the first time, LTER sites had the ability to acquire and use geo-referenced information systems. The current transformation derives from affordability: for the first time, generally available (and affordable) systems and tools make it possible to imagine an information-everywhere environment for ecoinformatics.

Twenty-five years ago, technology costs were a limiting factor. For an LTER site to deploy an advanced GIS capability, powerful Sun workstations each costing $20,000 had to be acquired for every person who wished to interact with the GIS. Staff costs were cheap, relative to the technology. Today, Moore's Law has changed things dramatically. Those "powerful" $20,000 workstations had less CPU power, less RAM, and less disk space than today's $500 iPad.

Now the challenge is to devise technical solutions that minimize manual operations by paid staff so that labor costs can be afforded. The proliferation of self-service devices (*e.g.*, airline check-in kiosks) are examples of this trend. In field ecology an equivalent example would be the development of data-acquisition systems that also automatically acquire the necessary metadata.

In the past, if a photograph of a study plot were to be used as data, additional metadata (date, time, and location) would have to have been manually recorded and associated with the photograph, and with all copies of the photograph. Today, most digital cameras automatically record date and time and embed the information into the image file itself, using the exchangeable image file format (EXIF) specification. Some cameras are also capable of detecting and recording location information using GPS information.

Clearly, in the long run it is much less expensive to acquire a more expensive data-capture system that automates *all* aspects of data and metadata capture than it is to use a less expensive instrument with attendant needs for manual processes.

Recommendation 1: LTER data-management operations should be optimized to take advantage of falling technology costs, especially in the area of automated data and metadata collection, while simultaneously maximizing the efficiency of LTER data-management staff.

Ensuring the Long-Term Availability of Long-Term Data

Since the beginning of LTER, three sites have dropped from the program — North Inlet, Illinois Rivers, and Okefenokee. As part of this review, I attempted to locate data from these three sites. I succeeded in locating data from North Inlet (both within the LTERNET system and on servers at South Carolina), but I was unable to locate *any* LTER data sets from either the Illinois Rivers or the Okefenokee sites.

Although both sites failed before there was any network-wide data archiving capabilities, the result is still ominous and demands attention. In two out of three cases on record, the

high finance) and then on through other fields where affordability becomes more important. With commodity desktop computers now more powerful than department-sized machines of thirty years ago, IT transformations are moving throughout society. In many areas, the last essential innovation is devising ways to minimize the labor costs associated with accomplishing the goal. This can involve automation or crowd-sourcing.

demise of an LTER site has apparently led to the total loss of LTER data sets associated with the defunct site.

From the perspective of LTER@100, it is imperative that LTER data outlive individual PIs, outlive individual sites, and even outlive individual research institutions.

The findings with Illinois Rivers and Okefenokee suggest that allowing a site that generates LTER data to be the prime (and only) site responsible for archiving and sharing those data is unacceptably risky.

Recommendation 2: NSF must amend the LTER funding guidelines to require that archiving and sharing of LTER data sets occur at third-party institutions, not at the site that creates the data. When evaluating a site's compliance with LTER data-sharing and data-access policies, only data hosted and made available through third-party sites should be considered shared. Any data hosted only at the producing site should not be counted among the site's shared data sets.

Economies of Scale

Achieving Economic Efficiency

The greatest improvements in the productive powers of labor, and the greater part of the skill, dexterity, and judgment, with which it is anywhere directed, or applied, seem to have been the effects of the division of labor.

Adam Smith, *An Inquiry into the Nature and Causes of the Wealth of Nations* (1776).

Collectively, the LTER sites employ more than 40 data-management FTEs. Working together, forty IT staff would constitute a fairly substantial IT department. Despite these aggregate resources, however, there are still many unmet data-management needs in LTER. Our review suggests that even doubling local site staffing to 3.0 FTEs per site would likely leave needs unmet, even though the resulting 80 FTEs would represent a larger IT staff than is typically found at a biomedical research facility with a thousand or more employees and an annual research budget well in excess of $150 million.

How can it be that such a large investment in IT staff yields the perception of such a relatively small result? *The problem lies not in the staff themselves, whom I have consistently found to be exceptionally competent, hard-working individuals.* Instead, the problem results from the staff being spread so thin that most individual LTER data managers must perform nearly the full breadth of data-management support activities. Economies of scale, optimization through specialization, cannot occur if staff are divided into small, self-contained workgroups of just a few individuals each. This is true no matter how skilled or dedicated the staff.

At the beginning of the industrial revolution, Adam Smith recognized that specialization is the key to economic efficiency and thus to economic growth. An economy in which individuals (or families) are responsible for meeting nearly all of their own needs (food, clothing, housing) is a subsistence economy in which few surpluses exist and individual wealth creation is minimal or non-existent.

To the extent that each LTER site's information needs are being met by the activities of one, or one and a fraction, data-management FTEs, that LTER site is operating in an *information subsistence economy*, with all of the non-economies of scale that that entails. Without effective interaction, the aggregation of 26 subsistence economies is just a larger subsistence economy. Although interactions among site data managers have resulted in substantial LTER-wide improvements over a pure subsistence economy, many significant problems associated with achieving economies of scale remain.

This problem has been recognized by many of the data managers. The advantages that occur when IT staff work together in a group, rather than always as individuals, is why one data manager told us that attending the annual data managers meetings is the most valuable event of the year and why others have banded together and pooled their data-management supplements to achieve some semblance of an economy of scale.

This problem was also recognized by the 20-year review committee, when they recommended that "NSF should support the informatics core function at a level sufficient to achieve the LTER program's informatics objectives in an aggressive timeframe," and then went on to note that this could be accomplished more efficiently by funding "an informatics hub, namely, a center that would be a proactive force in developing and promoting new informatics tools and technologies, with strong links to [relevant outside] groups" and then commenting explicitly that "concentration of informatics resources in a single hub would also achieve greater visibility and *economies of scale* than distribution of resources among existing sites."

Such an approach is not without risk, as is well known to anyone who has managed a central IT department in a research organization. The relationship between a central IT operation and localized research activities can quickly devolve into a dysfunctional state, unless the central unit is managed with a deliberate focus upon *providing services that meet the specialized needs of researchers.* It is possible to devise IT architectures that provide standard solutions to non-standard problems if sufficient care is taken to understand and then abstract the needs of individual researchers.

Many scientists have personal experience with non-responsive centralized IT departments and so may have a strong, and justified, reluctance to have any of their personal IT support move to a more centralized delivery mechanism, even if substantial economies are likely to be achieved. Therefore, efforts to mandate the use of centralized services can result in substantial resistance. It is often better to develop and deploy high-quality services first, without mandating their use. Many researchers who would resent being forced to use centralized services will voluntarily adopt them, if they provide cost-effective real solutions, and if they are responsive to changing research needs.

Recommendation 3: NSF and LTER must take steps to increase the efficiency of LTER data management by achieving economies of scale, while also being highly responsive to the needs of individual researchers.

The current level of staffing at typical LTER sites — less than 1.5 FTEs — probably represents the minimum possible, if real local needs are to be effectively met. Thus,

efficiencies cannot be achieved by cutting the local site staff, but only by augmenting a centralized staff at some *informatics hub*, as envisioned in the LTER 20-year review.

Activities at this hub could increase the efficiency and effective of both the overall LTER system and the individual LTER sites, if a substantial part of its effort were dedicated to providing tools that met many of the cyberinfrastructure needs of local LTER environments, while allowing local LTER data-management staff to focus on addressing scientific issues associated with local data management. One of the local LTER site data managers captured this in a comment, "What I'd really like to have is for someone to provide me with IM (information management) in a box."

This is not a fanciful request. Recent advances in virtualization and in the development of virtual servers as information appliances provide many opportunities for an LTER informatics hub to deliver hugely valuable tools to local LTER sites.

Recommendation 4: NSF and LTER must recognize that the most efficient way to meet LTER data-management needs is through the centralized acquisition and development of resources that can be effectively and efficiently made available to all LTER sites as services that can then be tuned by local data-management staff to meet local scientific needs..

Whether the development of an informatics hub is best done through a competition for a wholly new resource, or through the growth of the LTER Network Office is a tactical decision best left to NSF and LTER. I do not have access to enough LTER operational details to make recommendations at this level.

However, the review committee does note that the current work underway at the LNO with ARRA funding represents a very good first step in the direction of developing an LTER informatics hub. Whatever the long term decisions regarding the creation of a hub, it is important that the one-time ARRA funding not be allowed to create a burst of centralized IT activity that then ends abruptly when the ARRA funding expires.

Recommendation 5: NSF and LTER should work together to avoid a hard landing at the end of the ARRA funding for cyberinfrastructure.

While it is clear that additional investment in LTER data-management infrastructure will be needed if the LTER@100 vision is to be achieved, it is equally clear that any additional investment to improve long-term data management at LTER *must* be done in a leveraged manner. The single worst way to do it would be to spend the same amount at each LTER site. Spending money evenly would spread it too thin, resulting in the lose-lose combination of high aggregate price and low overall effectiveness.

Figure 6 shows the relative budget and size of the overall LTER network, relative to the budget for the LTER Network Office (LNO). Through the year 2000, the budget of the LNO tracked the overall growth of the LTER program, but then for nine years the LNO budget flattened, while the overall LTER program continued to grow. From the perspective of achieving economies of scale in data management, this is ill-advised.

As the figure shows, it is only with the burst of ARRA funding that the budget of the network office regained parity with the overall LTER program. From the perspective of data-management efficiency, increasing local site-specific spending while decreasing resources available to an informatics hub is precisely the wrong thing to do.

Recommendation 6: If economies of scale are to be achieved, growth on spending for a centralized informatics hub must exceed growth in spending on the overall LTER network until such point as a functioning hub is established.

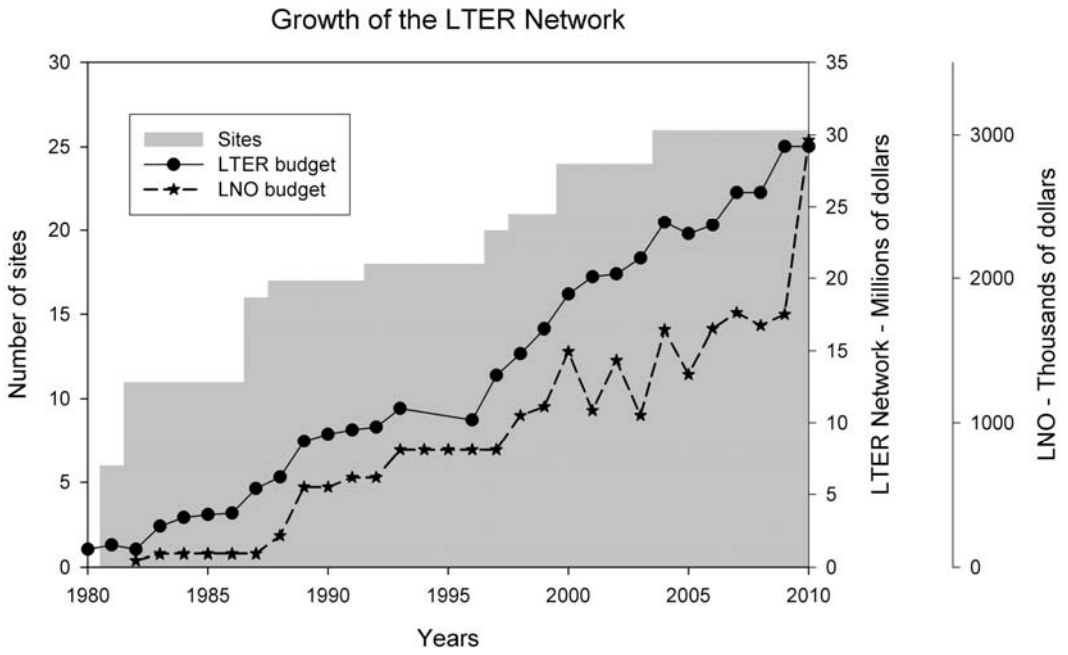


Figure 6. A comparison of the relative investment in the LTER Network Office and the in the overall LTER program. The figure is taken from *The LTER Network Office: An Overview of Recent Developments,* a presentation given to the LTER 30-year Review Committee by Robert B. Waide, Executive Director, LTER Network Office.

Leveraging an Informatics Hub

Several opportunities exist for NSF and LTER to leverage an informatics hub to great overall gain. Such a hub could provide IT services from the hub directly to the site-specific IT staff (*i.e.*, provide the "IM in a box" requested by one data manager). For example, the hub could implement a large virtual server farm and offer virtual server hosting as a service to individual LTER sites.[23] To make all of this work, the hub should have a dedicated employee serving as a *customer-account manager* to oversee the relationship between the hub and site-specific IT activities. If this task is performed well,

[23] In a keynote address to a recent national Bio-IT national meeting, Chris Dagdigian described the development of such vCOLO (virtual co-location) facilities to provide centralized server support for decentralized IT departments as *one of the greatest success stories of the year*.

local IT and scientific leaders would see the customer-account manager as their person inside the hub, looking out for their interests.

Services to the local LTER sites should be based on formal service-level agreements, with clearly spelled out service and quality levels. Additional leverage could be accomplished by funding a few local site-specific data managers to serve on the hub staff, on a rotating basis. Current communication capabilities could allow that to happen effectively in a purely virtual mode, without requiring relocation to Albuquerque or wherever the hub is located. Having teams of rotating local data managers acting as staff to the informatics hub would help ensure a meaningful connection between the hub and the needs of local sites (while also helping local data managers appreciate the difficulties and challenges of providing centralized services).

To deliver maximum value, the informatics hub should be run as a mature informatics service facility, always striving for the highest possible level of operational maturity. Appendix III describes one approach to measuring the maturity of IT operations. Historically, many scientific IT departments have been characterized by the peak-performance heroics of level I, sometimes moving into Level II. To deliver the kind of high-performance, yet sustainable infrastructure needed by LTER@100, the IT activities associated with LTER must make significant movement up the maturity hierarchy.[24]

Data Publication, not Data Sharing

The process of LTER *data sharing* needs to be rethought into a model of *data publishing*, with defined data products and services. So long as access to LTER data is through individual, idiosyncratic, site-specific web sites, so long will LTER data be at risk and accessing LTER data be tedious and frustrating. Shifting to a data publishing model will not, to be sure, magically solve all problems, but it will help to control expectations, to facilitate standardized search and access, and to encourage the development of third-party tools to assist in the use of the published data. It will also allow the development of formal specifications regarding the published data objects, thus providing an answer to the question, how much will be enough? (*cf.* the discussion on page 5).

Recommendation 7: LTER sites and researchers should be required to develop their data sets into data products, then share those products by submitting them to approved third-party data publishers.

Of course, this begs the question of who, exactly, will provide these data publishing services. In 2003, the Atkins Report explicitly called for the creation of central data repositories and current movement towards NEON and DataONE, not to mention the possible development of an LTER informatics hub, all suggest that such facilities are, or soon will be, available. The optimum future service model envisioned by Peters and illustrated in Figure 3 (page 6, above) depends upon such repositories and it is difficult to image how the full goals of LTER@100 could be achieved without them.

---

[24] The need for maturation in scientific computing support is recognized as a general problem, not specific to LTER. In 2006 a national meeting of BRIITE (Biomedical Research Institutions Information Technology Exchange) was dedicated to the need for increased maturity in scientific computing infrastructure.

http://www.esp.org/briite/meetings/2006/MSK/index.html

Manipulating Incentives, Shaping Behavior

The manipulation of incentives to shape behavior must be done, at least in part, by funding agencies and publishers. Only they have the ability to control the two most important incentives (access to research funds and to publishing outlets) that affect the behavior of most academic scientists. If the past history from other communities is any guide, good behavior does not spontaneously emerge from some sort of community maturation. It occurs as the result of active social engineering carried out by visionary leaders with access to critical handles of power — the ability to modify incentives, including both rewarding those who share data appropriately and punishing those who "steal" protected data as well as those who horde "public' data.[25]

NSF could simply put in place more stringent requirements for the recipients of LTER funding, such as requiring annual PI reports to document the provisioning of data for others or in holding interim reviews of LTER sites to assess their success in meeting the LTER@100 needs for data sharing and publication.

When it comes to infrastructure deployment, funding agencies should not try to manage infrastructure procurement the same way they manage the funding of research. For LTER to reach its full LTER@100 potential, NSF leadership and staff need to be engaged, especially regarding LTER activities not related to site-specific excellence.

Recommendation 8: NSF leaders and staff must become more actively engaged in the planning, management, and oversight of LTER operations, especially with regard to long-term activities that must ultimately transcend the interests of any one site or any one PI.

Creating long-term data sets for others to use is a service function, and scientific careers are not built upon service functions. It is unreasonable to expect that local LTER PIs will somehow "rise to the occasion" and get this done. Spontaneous solutions have not happened in other scientific communities and they likely won't happen in LTER.

NSF operates at the juncture between the worlds of research and of politics. Funds are involuntarily extracted from taxpayers, then politically allocated to funding agencies to support research in the hopes of ultimately returning enough public value to justify the expenditure. Funding agency staff then oversee an award process that is intended to maximize the quality of science performed in order to maximize the political return on the original allocation.

If, as some believe, that LTER will return its greatest value over the longest possible time scale, then managing the current activities in support of potential future return is as much a political process as a scientific one, and only agency staff, not field scientists, are

[25] It is important to bear in mind that the goal of data sharing must be matched with respecting the rights of those who create and collect data. NSF achieves its mission best when its researchers are fully dedicated to their research. Data sharing, if done inappropriately, threatens this dedication. What if an individual who created an experiment, gathered the data from it, assembled them into a data base, was then 'scooped' by someone who had time available to more quickly publish the results. It would be a rare scholar who would not be disheartened buy such an event, and feel that the other scientist had cheated. As such, data sharing rules must protect the personal investment of all investigators while also allowing others to have access to the data within a reasonable amount of time.

positioned to function at the interface of science and politics, that is, to play the role of politically informed scientific leader.[26]

Understanding and Stimulating Usage

To date, the third-party use of LTER data sets has been relatively light, with much use going for educational purposes. This has likely been due, in part, to the fact that LTER is still just getting started (thirty years isn't that long when it comes to assessing truly long-term phenomena), and also in part to the fact that accessing and using the data is still a non-trivial task.

Until now, NSF has focused largely on funding the supply side of long-term ecological data. I suggest that NSF also consider funding the demand side, either through special competitions or special supplements or even one-time contests. In addition, NSF and LTER would be well advised to take active steps to understand both the demand for long-term data and the structural and metadata constraints that must be placed on long-term data to make them truly useful.

Understanding how to collect data so that they may be arbitrarily combined, yet still yield good science, is a scientific problem, not a technical one. This problem could be investigated through workshops or special meetings, such as the catalysis meetings occasionally held at NESCent.

Additionally, LTER, either through the current network office or through a future informatics hub (should one be created), could periodically convene focus groups of scientists who have downloaded LTER data. The real needs of third-party users can only be appreciated by interacting with third-party users. LTER scientific and technical staff are too close to the LTER program itself to fully appreciate the potential, and the problems, associated with the use of LTER data for non-LTER purposes.

Recommendation 9: Both NSF and LTER should take steps to better understand the needs of potential users of LTER data.

Interesting and relatively inexpensive efforts could also be made to stimulate the use of LTER and the development of effective presentation modes for LTER data. For example, Hans Rosling (a Swedish statistician) has devised methods for presenting multi-dimensional time-series data in a way that is certainly effective and often exciting (see Appendix II).

Funding research into the presentation of long-term data could be expensive and time consuming. However, offering a relatively modest annual prize for the best use and

[26] When NSF started the Arabidopsis genome project, NSF insisted on the program including an international database of findings and one or more stock centers to house and distribute strains. At the time, many members of the scientific community considered this bureaucratic micro-management and felt that the money would be better spent directly on research. However, a few years into the program, the value of the infrastructure became more evident. At one meeting, a prominent Arabidopsis researcher — Elliot Meyerowitz — went out of his way to tell Machi Dilworth that NSF was right and that more total research got done because of the investment in infrastructure than could have gotten done if ALL of the budget had gone to science, and he thanked her for NSF's vision in seeing what it would take to succeed. Similar vision and leadership will be needed to deliver LTER@100.

presentation of LTER data could build interest and yield results quickly. For example, NSF (or LTER) could offer a small annual prize for the most effective presentation created and published (on the web, as a video) using LTER data. A $10,000-dollar prize would likely engage the efforts of many graduate students and post docs, while a $50,000 prize would draw in an even bigger community, including scientists and graphic designers. Even at $50k, the cost would be minimal if it resulted in more awareness of the data and their availability, and if it helped inform the LTER community about how best to format and serve the data for others to use.

Recommendation 10: NSF and LTER should take steps active steps to encourage the effective, creative, and regular use and presentation of LTER data.

Work with Other Programs and Services

LTER cannot function effectively in a vacuum, now or in the future. Every effort should be made to ensure continued and new interactions between LTER and other relevant activities such as NEON, NCEAS, NESCent, and DataONE.

From a truly long-term perspective, the value of LTER data should even outlive the LTER program itself. This can only happen, if LTER work, findings, and data are effectively embedded in a larger community.

Recommendation 11: NSF and LTER should actively support collaborations and interactions between LTER and other relevant activities. To be especially effective, such support should be available on fairly short notice, with minimal procedural impediments. To encourage spontaneous interactions, some form of after-the-fact awards or prizes might be considered. The advantage of after-the-fact awards is that they stimulate activity before payment is made and ultimately payment need only be made for those activities deemed sufficiently worthy.

Although a contest or prize mechanism is not feasible for funding actual research, with fairly large attendant budgets, it can be a very effective manner of funding smaller, short-term projects. Indeed, such a mechanism has even been effectively employed to allow the crowd-sourcing of technical problems in industry.

Realistic Expectations and Realistic Resources

The 20-year review committee called out the importance of tying LTER goals and objectives to realistic budgets. This is especially important in the area of data sharing, where expectations tend to be unrealistically high. It will be helpful if NSF and LTER can work together to move from a vague notion of data sharing to a more defined notion of data publishing. With data publishing, specific types of data objects can be defined to optimize the tradeoffs between ease of use and cost of creation.

Then, if LTER and NSF jointly agree on the specifications of the data products to be produced it will be readily apparent whether or not LTER is delivering on its commitments or if members of the research community have expectations at variance with what has been promised. By working together with community user groups to understand and assess the needs of potential data users (*cf.* Recommendation 9) it will be possible for both NSF and LTER to decide what is practicable and how much is enough.

To avoid unnecessary frustration, it is important that these determinations be documented and made generally available, so that members of the user community can know what they can reasonably expect. For example, GenBank has long taken great pains to document what aspects of the data it manages (the sequence itself) can be considered primary and therefore will be maintained in a stable format forever, and what aspects are considered secondary (commentary on the sequences, including the identification of genes) and thus may be subject to undocumented format changes.

In 2003 NSF released the Atkins Cyberinfrastructure report[27] in which a blue-ribbon panel noted that "We now have the opportunity and responsibility to integrate and extend the products of the digital revolution to serve the next generation of science and engineering research and education" and called upon NSF to recognize that:

> Achieving the vision of the Advanced Cyberinfrastructure Program (ACP) will require coordinated NSF support of a broader set of activities and facilities than the agency has historically supported. In addition, existing activities (e.g. providing access to high-end computers, enduring data archives, and middleware software development) will need substantially higher funding levels.

In particular, the report recommended that "NSF, in collaboration with other appropriate mission agencies, should take lead responsibility for creating and maintaining the crucial data repositories necessary for contemporary, data driven science." The report estimated that adequate data repositories would cost on the order of $185 million per year and explicitly noted that "These amounts are meant to be in addition to the current NSF investments in these areas."

In discussing details of the envisioned data repositories, the report asserted, "To illustrate some detailed issues, data need to be organized in appropriate ways, metadata (machine readable and searchable descriptions of the data) must be systematically created, and basic manipulation and analysis tools provided."

Although federal budget realities did not allow NSF to implement an Advanced Cyberinfrastructure Program on the scale envisioned in the report,[28] data-management work done in the LTER community has closely followed the directions outlined in the report, especially in the area of metadata and tool development.

Along with offering an inspiring vision for cyberinfrastructure-enabled science, the Atkins report noted the need for breath-taking expenses to implement that vision: more than a billion dollars per year in new spending. As NSF and the LTER community lay out

---

[27] *Revolutionizing Science and Engineering Through Cyberinfrastructure: Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure* http://www.nsf.gov/od/oci/reports/atkins.pdf

[28] Who in 2002 (when the report was being written) could have foreseen the coming trillion-dollar expense of the wars in Iraq and Afghanistan and the near economic meltdown that would begin in 2007 and which is not yet fully resolved?

strategic plans for implementing the full vision of LTER@100, it is important that they accompany that vision with appropriate plans for resource allocation.

Recommendation 12: In crafting strategic plans, NSF and the LTER program must tie the scientific goals and objectives to a realistic budget. NSF should allocate funding for the LTER program commensurate with the agreed goals and priorities for synthesis science, ecological forecasting, and data management in the fourth decade.

REFERENCES:

Bruna, Emilio M. 2010. Scientific Journals Can Advance Tropical Biology and Conservation by Requiring Data Archiving. *Biotropica* 42:399-401.

Costello, M J. 2009. Motivating Online Publication of Data. *Bioscience* 59: 418-427.

Hawking, Stephen W. 1988. *A Brief History of Time*. New York: Bantam Books.

LaBare, K M, R L Klotz, and E Witherow. 2000. Using Online Databases to Teach Ecological Concepts. *The American Biology Teacher* 62: 124-127.

Magnuson, John J. 1990. Long term Ecological Research and the Invisible Present. *Bioscience* 40: 495-501.

Michener, W. K., Brunt, J. W., and S. G. Stafford. 1994. *Environmental Information Management and Analysis: Ecosystem to Global Scales*. London: Taylor & Francis LTD.

Michener, W K, J W Brunt, J J Helly, T B Kirchner, and S G Stafford. 1997. Nongeospatial Metadata for the Ecological Sciences. *Ecological Applications.* 7: 330-342.

Moore, Allen J, Mark A McPeek, Mark D Rausher, Loren Rieseberg, And Michael C Whitlock. 2010. The Need for Archiving Data in Evolutionary Biology. *Journal of Evolutionary Biology.* 23:659-660.

Peters, Debra P C. 2010. Accessible Ecology: Synthesis of the Long, Deep, and Broad. *Trends in Ecology & Evolution.* 25: 592-601.

Porter, J H. 2010. A Brief History of Data Sharing in the US Long Term Ecological Research Network. *Bulletin of the Ecological Society of America* 91: 14-20.

Porter, J. H., and J. T. Callahan. 1994. Circumventing a dilemma: Historical approaches to data sharing in ecological research. In Michener et al. [eds.] *Environmental Information Management and Analysis: Ecosystem to Global Scales*. London: Taylor & Francis LTD.

Rausher, Mark D, Mark A McPeek, Allen J Moore, Loren Rieseberg, and Michael C Whitlock. 2010. Data Archiving. *Evolution.* 64:603-604.

Whitlock, Michael C. 2010. Data Archiving in Ecology and Evolution: Best Practices. *Trends in Ecology & Evolution.* 26: 61-65.

APPENDIX I: RECOMMENDATIONS FROM LTER 20-YEAR REVIEW

Recommendation 1. The committee concurs and recommends that the LTER program forge a bold decade of synthesis science, one that will lead to a better understanding of complex environmental problems and result in knowledge that serves science and society.

Recommendation 2. In order to achieve the full promise of synthesis science, the LTER program should adopt and make systemic what NSF has informally termed "21st century biology", namely, that LTER science be multidisciplinary, multidimensional, scalable, information driven, predictive and model based, education oriented, and increasingly virtual and global.

Recommendation 3. If the third LTER decade is to be one of synthesis science, the LTER program must define its niche, namely, it needs to determine its priorities for synthesis science and what the scientific focus or foci of such synthesis will be.

Recommendation 4. Ecological research by LTER scientists involving multiple disciplines, dimensions and scales should be organized *a priori* by hypotheses and theory, and tested by predictive models across broader and broader phenomena.

Recommendation 5. The LTER community should review the role, theoretical basis, scope, function and continued usefulness of the core areas, especially in formulating its priorities for synthesis science in the next decade.

Recommendation 6. The committee recommends that biological diversity be designated a new core area (or function) for the LTER program at all or selected sites and receive appropriate levels of funding.

Recommendation 7. The LTER program should become a research collaboratory a seamless, integrated continuum from site-specific to cross-site to network-wide and systems-level ecological research. Building on its successes to date, the LTER program should become more collaborative across ecological and other research communities. To do so, it must increase its recruitment of scientists, technologists and expertise from outside traditional LTER disciplines who will help formulate hypotheses and apply technologies that will advance ecological science.

Recommendation 8. NSF should provide real incentives primarily funding and competitions on a regular basis to encourage cross-site, interdisciplinary, systems-level collaboration to enhance theory, reveal large-scale ecological phenomena and inform environmental policy.

Recommendation 9. NSF should revise LTER and LTER-related proposal guidelines and review criteria to provide greater balance and synthesis between site-specific and cross-site research and education. In doing so, NSF should consider placing site-specific and cross-site competitions and activities on parallel, complementary tracks.

Recommendation 10. The LTER program should establish informatics as a *core function* by implementing a systemic informatics infrastructure and architecture that integrates LTER data and tools with those from relevant disciplines.

Recommendation 11. NSF should: (1) increase the importance of data management and informatics in its evaluations of LTER activities and in its requests for proposals,

consistent with the importance of these issues in LTER's coming decade of synthesis; (2) support the informatics core function at a level sufficient to achieve the LTER program's informatics objectives in an aggressive timeframe.

Recommendation 12. The LTER community should aggressively incorporate powerful new scientific approaches, technologies and analytical and experimental tools that can expand the scope and scale of LTER science to systems-level ecological research. In doing so, the LTER program must identify and select the disciplines, approaches and technologies appropriate to achieving its scientific priorities and agenda for its decade of synthesis. For its part, NSF must boost funding levels to enable incorporation of new science and technologies in order to enhance the scale, scope and tractability of LTER's ecological research.

Recommendation 13. The LTER program should partner with social scientists to increase understanding of the interrelationships and reciprocal impacts of natural ecosystems and human systems in order to inform environmental policy.

Recommendation 14. The LTER program should foster increased opportunities for serendipitous science by providing a virtual portal to its legacy data for investigators worldwide.

Recommendation 15. Using the knowledge gained from synthesis and serendipitous science in the coming decade, the LTER program should assume a more powerful and pervasive role in informing environmental solutions and public policy at local, national and international levels. The LTER program should increase and regularize efforts to share this wealth of knowledge with public officials, especially at the national level where environmental policies can have significant impact.

Recommendation 16. The LTER program should establish a professional public communications office to assist LTER scientists in informing the public and policy makers about the importance of LTER science to local, regional and national environmental solutions. NSF should provide a budgetary line item for this function.

Recommendation 17. The LTER program should expand the scope of its undergraduate and graduate education in field-based ecological research by incorporating the cross-disciplinary, collaborative approaches and characteristics of 21st century biology.

Recommendation 18. The implementation and impact of the Schoolyard LTER should be enhanced in three ways. First, the LTER sites should leverage funds provided for this program to achieve economies of scale and increased outreach; second, NSF should increase its support for LTER K-12 educational programs; and third, the design and outcomes of LTER K-12 educational programs should have formal evaluation to inform appropriate growth and improvements.

Recommendation 19. The LTER community and NSF, using the findings and recommendations of this report, should jointly craft and implement a comprehensive strategic plan for the LTER program, i.e., its science, funding, outcomes, governance and organization for the next decade. The comprehensive strategic plan should contain all the components basic to any strategic plan: vision and mission; goals, priorities, objectives and actions; deliverables, timelines and milestones; and a budget that aligns resources with these elements.

Recommendation 20. In crafting the strategic plan, NSF and LTER program must tie the scientific goals and objectives to a realistic budget. NSF should increase funding for the LTER program commensurate with the agreed goals and priorities for synthesis science and ecological forecasting in the third decade.

Recommendation 21. The enhanced budget for the LTER program should be invested in the LTER's scientific priorities and in implementing the strategies discussed in this report for achieving 21st century biology and synthesis science.

Recommendation 22. The NSF should establish parity funding for all LTER sites as quickly as possible commensurate with the scientific goals and activities called for by individual sites and the network in the strategic plan.

Recommendation 23. New sites should not be added to the LTER network until such potential expansion is justified in the strategic plan.

Recommendation 24. Should sites be added to the LTER program in the future, such expansion must be strategic and synoptic. The larger ecological community (i.e., LTER and non-LTER ecological communities) should determine where and how such expansion would provide the greatest benefit to understanding the nation's ecological systems, with competitions for new sites based on these findings. The LTER program should expand internationally by building on its collaborations with the ILTER enterprise.

Recommendation 25. The comprehensive strategic plan should describe a governance and organizational structure appropriate to the goals, scope and scale of the LTER program in the next decade of synthesis science. Planning this structure should be informed by models from other enterprises and by experts in academia, government and the private sector.

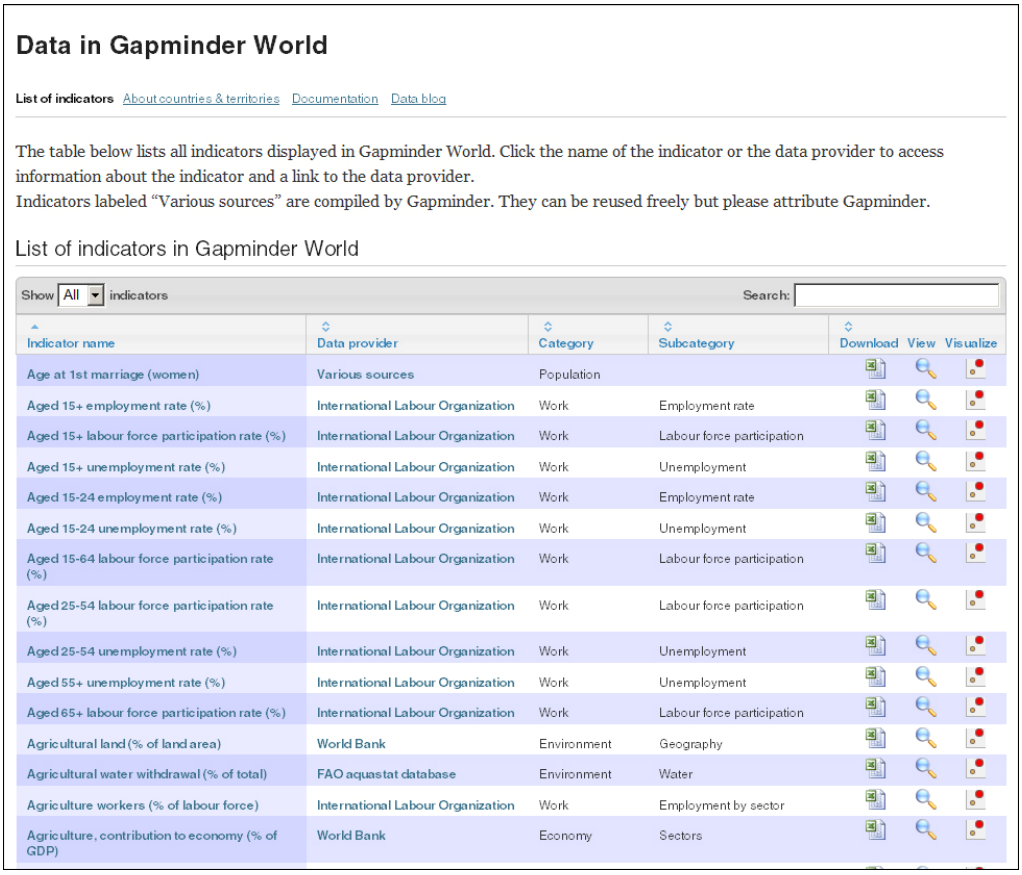
Recommendation 26. The strategic plan should specify the entity or entities that will implement the strategic plan and manage the LTER program, as well as a process for developing policies to govern implementation, LTER management and other issues.

Recommendation 27. NSF should allocate two program officers to administer the LTER program, with a permanent one in charge of the program. Because LTER is now funded by several NSF directorates, a formal, cross-directorate committee of program officers should be established to coordinate LTER funding and program management.

APPENDIX II: EFFECTIVE PRESENTATION OF LONG-TERM DATA

Access to long-term data is helpful, but mere access still leaves a lot of work to the user before even a beginning appreciation of the significance of the data can be had. Han Rosling is a Swedish statistician who has developed an incredibly effective way of presenting long-term, time-series data, using custom software — GapMinder[29]. GapMinder presents three-dimensions of data in two dimensions, then adds time as a fourth dimension by presenting a series of two-dimensional time snapshots as a video. As an example, of the increase in effectiveness that occurs when one progresses from *locating* data, to *accessing* data, to *visualizing* data, consider the following.

Rosling's GapMinder site makes available a number of useful data sets, through an interface that looks a lot like any other data catalog:[30]



Data in Gapminder World

List of indicators About countries & territories Documentation Data blog

The table below lists all indicators displayed in Gapminder World. Click the name of the indicator or the data provider to access information about the indicator and a link to the data provider.
Indicators labeled "Various sources" are compiled by Gapminder. They can be reused freely but please attribute Gapminder.

List of indicators in Gapminder World

Show All indicators Search:

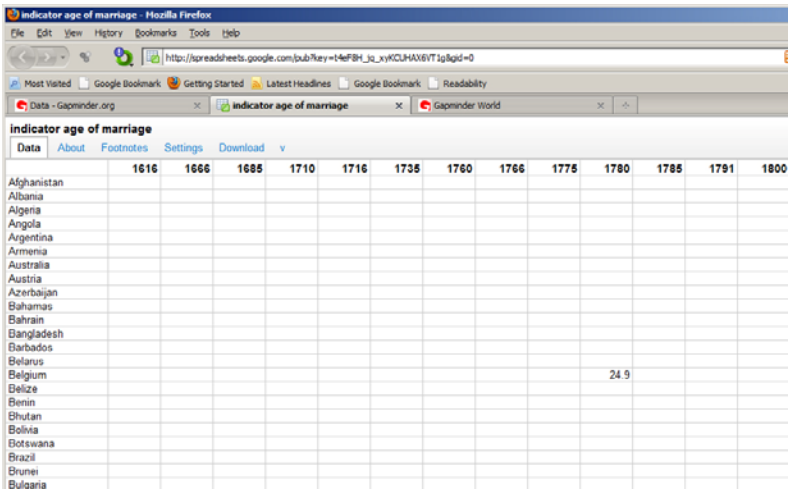
Indicator name	Data provider	Category	Subcategory	Download	View	Visualize
Age at 1st marriage (women)	Various sources	Population				
Aged 15+ employment rate (%)	International Labour Organization	Work	Employment rate			
Aged 15+ labour force participation rate (%)	International Labour Organization	Work	Labour force participation			
Aged 15+ unemployment rate (%)	International Labour Organization	Work	Unemployment			
Aged 15-24 employment rate (%)	International Labour Organization	Work	Employment rate			
Aged 15-24 unemployment rate (%)	International Labour Organization	Work	Unemployment			
Aged 15-64 labour force participation rate (%)	International Labour Organization	Work	Labour force participation			
Aged 25-54 labour force participation rate (%)	International Labour Organization	Work	Labour force participation			
Aged 25-54 unemployment rate (%)	International Labour Organization	Work	Unemployment			
Aged 55+ unemployment rate (%)	International Labour Organization	Work	Unemployment			
Aged 65+ labour force participation rate (%)	International Labour Organization	Work	Labour force participation			
Agricultural land (% of land area)	World Bank	Environment	Geography			
Agricultural water withdrawal (% of total)	FAO aquastat database	Environment	Water			
Agriculture workers (% of labour force)	International Labour Organization	Work	Employment by sector			
Agriculture, contribution to economy (% of GDP)	World Bank	Economy	Sectors			

The right-hand side offers the user the opportunity to *download* the data (as an Excel file), to *view* the data (through the web interface), and to *visualize* the data (through the GapMinder animation tool).

If one clicks *view* on the first data set, one gets a typical view into a two dimensional (sparsely populated) set of data showing age at first marriage for women, by country by year.

29 http://www.gapminder.org/

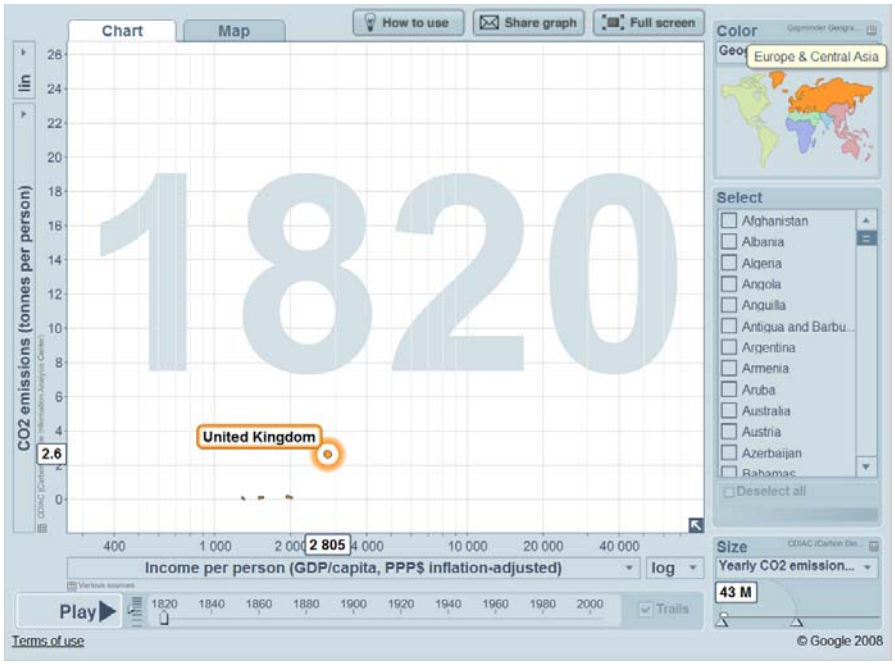
30 http://www.gapminder.org/data/



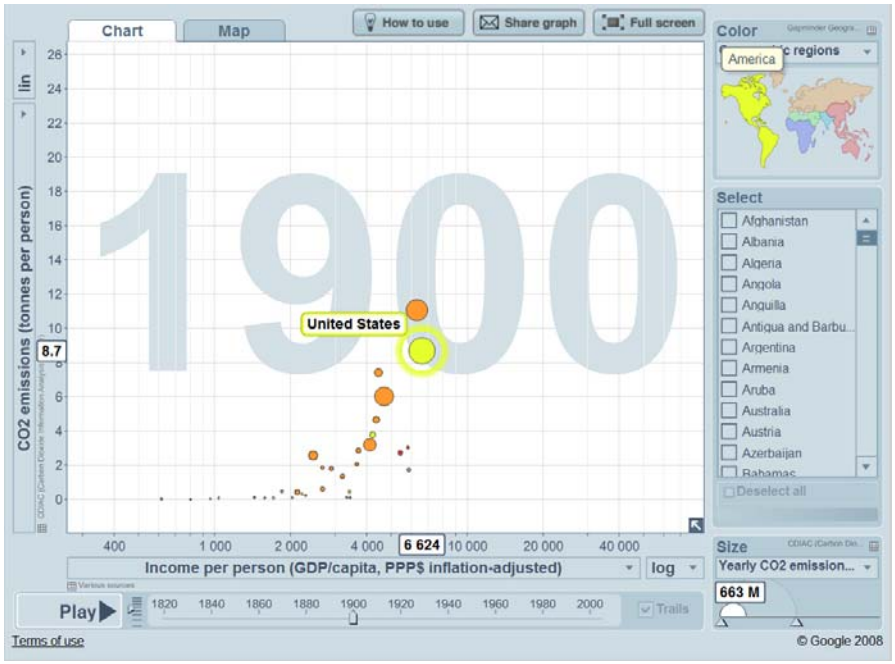
If one clicks visualize, one is presented with a display that is easy to understand, even though substantially more data have been included, including per-capita income, and total population size of the country (size of the circle that represents the country).

There appears to be a distinct correlation between increasing income and increasing age at first marriage. But, is that a simple causal correlation or are other forces at work? By clicking play (button in the lower left corner), one can watch the data change over time, from 1800 to the present. During the 19th century, one sees only left-to-right movement, indicating that changes in average income did not affect age of first marriage for women. In most of the 20th century, one sees generally rising ages of first marriage, but without a pronounced connection to increasing income. However, from 1975 on, the relationship of age-of-first-marriage with income begins to appear, but the most pronounced effects occur only after 2000. All of that information can be presented in a video that takes less than 45 seconds to play.

Although the GapMinder site is largely concerned with national statistics, such as health care, population size, and income, the tool could easily be used to present data of interest to ecologists. For example, the site does have data sets dealing with the production of CO_2 emissions over time.

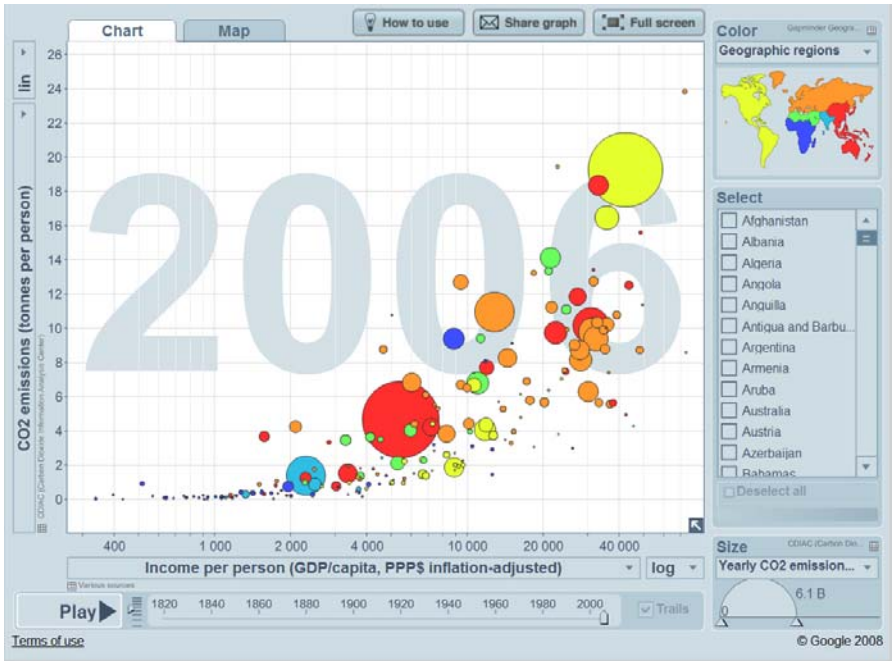


In 1820, at the beginning of the industrial revolution, the UK was the largest producer of CO_2 emissions, both on a per-capita basis (y-axis) and as a national total (size of the dot representing the UK).



By 1900, the United States had emerged at the largest aggregate producer of CO_2, even though on a per-capita basis it was still behind the UK.

At the end of World War II, the United States was clearly the world's dominant producer of CO_2, both on an aggregate and a per-capita basis.



In 2006, however, Luxembourg (the small orange dot in the upper right) had become the world's largest per-capita producer of CO_2, while China (the large red dot) was the largest aggregate producer of CO_2.

This “GapMinder” style of presentation could easily and effectively be used to present some kinds of long-term ecological data. For example, imagine such a presentation of a multi-decadal set of data relating the population size of some aquatic species in different lakes (one dot per lake, size of the dot corresponding to population size), plotted against, say, average pH and temperature of the lake, over time.

APPENDIX III: THE CAPABILITY MATURITY MODEL

The capability maturity model (CMM) was developed by Carnegie Mellon for the Air Force as a method for judging the capabilities of software developers. The CMM model has five levels:

Maturity Level 1: Initial

At maturity level 1, processes are usually ad hoc and the organization usually does not provide a stable environment. Success in these organizations depends on the competence and heroics of the people in the organization and not on the use of proven processes. In spite of this ad hoc, chaotic environment, maturity level 1 organizations often produce products and services that work; however, they frequently exceed the budget and schedule of their projects.

Maturity level 1 organizations are characterized by a tendency to over commit, abandon processes in the time of crisis, and not be able to repeat their past successes again.

Maturity Level 2: Repeatable

At maturity level 2, software development successes are repeatable. The organization may use some basic project management to track cost and schedule.

Process discipline helps ensure that existing practices are retained during times of stress. When these practices are in place, projects are performed and managed according to their documented plans.

Project status and the delivery of services are visible to management at defined points (for example, at major milestones and at the completion of major tasks).

Basic project management processes are established to track cost, schedule, and functionality. The necessary process discipline is in place to repeat earlier successes on projects with similar applications.

Maturity Level 3: Defined

At maturity level 3, processes are well characterized and understood, and are described in standards, procedures, tools, and methods.

The organization's set of standard processes is established and improved over time. These standard processes are used to establish consistency across the organization. Projects establish their defined processes by the organization's set of standard processes according to tailoring guidelines.

The organization's management establishes process objectives based on the organization's set of standard processes and ensures that these objectives are appropriately addressed.

A critical distinction between level 2 and level 3 is the scope of standards, process descriptions, and procedures. At level 2, the standards, process descriptions, and procedures may be quite different in each specific instance of the process (for example, on a particular project). At level 3, the standards, process descriptions,

and procedures for a project are tailored from the organization's set of standard processes to suit a particular project or organizational unit.

Maturity Level 4: Quantitatively Managed

Using precise measurements, management can effectively control the software development effort. In particular, management can identify ways to adjust and adapt the process to particular projects without measurable losses of quality or deviations from specifications.

Sub-processes are selected that significantly contribute to overall process performance. These selected sub-processes are controlled using statistical and other quantitative techniques.

A critical distinction between maturity level 3 and maturity level 4 is the predictability of process performance. At maturity level 4, the performance of processes is controlled using statistical and other quantitative techniques, and is quantitatively predictable. At maturity level 3, processes are only qualitatively predictable.

Maturity Level 5: Optimizing

Maturity level 5 focuses on continually improving process performance. Quantitative process-improvement objectives are established and used as criteria in managing improvement. The effects of deployed improvements are measured and evaluated against the objectives. Both the defined processes and the organization's set of standard processes are targets of measurable improvement activities.

Improvements to address common causes of variation and to improve the organization's processes are identified, evaluated, and deployed.

A critical distinction between maturity levels 4 and 5 is the type of process variation addressed. At level 4, processes are designed to address special causes of process variation and to provide statistical predictability of the results. Though processes may produce predictable results, the results may be insufficient to achieve the established objectives.

At level 5, processes are concerned with addressing common causes of process variation and with changing the process to improve performance (while maintaining statistical probability).

APPENDIX IV: ACCESSIBLE ECOLOGY

Peters, Debra P C. 2010. Accessible Ecology: Synthesis of the Long, Deep, and Broad. *Trends in Ecology & Evolution.* 25: 592-601.

Review

Special Issue: Long-term ecological research

Accessible ecology: synthesis of the long, deep, and broad

Debra P.C. Peters

Jornada Basin Long Term Ecological Research Program and USDA ARS, Jornada Experimental Range, Las Cruces, NM 88003, USA

Large volumes of data have been collected to document the many ways that ecological systems are responding to changing environmental drivers. A general buy-in on solutions to these problems can be reached only if these and future data are made easily accessible to and understood by a broad audience that includes the public, decision-makers, and other scientists. A developing framework for synthesis is reviewed that integrates three main strategies of ecological research (long-term studies; short-term, process-based studies; and broad-scale observations) with derived data products and additional sources of knowledge. This framework focuses on making data from multiple sources and disciplines easily understood by many, a prerequisite for finding synthetic solutions and predicting future dynamics in a changing world.

Challenges to synthesis

Dramatic changes in climate, land cover, and habitat availability have occurred over the past several centuries influencing every ecosystem on Earth [1,2]. Large amounts of data, and in particular observations over long time periods, have been collected to document changes, which include shifts in species dominance, loss of biodiversity, and reductions in clean air and water quality and quantity [3–5]. Solutions to environmental problems are elusive, in large part because much of the data have not been synthesized and remain inaccessible to a broad audience [6,7]. The complex nature of environmental problems requires that different types of data from multiple sources and disciplines be integrated [8], yet the sheer volume and nature of the data make it a challenge to ensure accessibility in a coherent, easy-to-understand format. Most data are too technical or complicated for general use [7], and many data are posted online in non-standard formats. Inaccuracies in the data and missing descriptive metadata further limit accessibility [9]. Some complex data have been distilled into useful formats for non-scientists [1,7], but questions can arise as to how the data were interpreted or analyzed (e.g. http://www.eenews.net/public/climatewire/2009/11/24/1). Standardization, simplification, and integration are required before data can be visualized, analyzed, and synthesized to generate new understanding [10,11].

Given that the Earth is changing at faster rates and in different ways than expected, there is a critical need to make existing and future data accessible in a format that the public and decision-makers can understand [12]. Accessible data are also needed by scientists to guide the strategic collection of additional data, and in synthesis efforts to yield new knowledge, insights, generalities, and solutions [8,13,14]. The continued collection of long-term data [15] and the emergence of observatories of multiple sites collecting a large suite of standardized data, such as the National Ecological Observatory Network (NEON), will magnify the problem further [16]; thus, reinforcing the critical need to improve data accessibility and utility within a synthesis framework that is sufficiently flexible, expandable, and robust to handle these future data sources.

Here, I review three general strategies associated with ecological research (long-term studies; short-term, pattern-process studies for deep understanding; and observation networks of sites for broad-scale patterns) commonly used to investigate ecological responses to a changing environment. Each strategy provides unique insights with important contributions to ecological knowledge, yet each also has scientific limitations and challenges to data accessibility and synthesis. Although examples of each strategy are drawn primarily from US-funded research, the principles and challenges apply globally [17,18]. Then, I describe a framework for synthesis being developed to make different types and sources of data from these strategies accessible with utility to a broad audience. I draw upon insights from the EcoTrends Project (http://www.ecotrends.info) to illustrate the application of this framework for a range of ecosystems found globally (terrestrial, aquatic, coastal, and urban) [19]. Finally, I emphasize new research directions to improve data accessibility and synthesis, and to provide new ecological knowledge for forecasting future ecosystem dynamics.

Ecology of the 'long'

The importance of long-term data to ecological knowledge has become increasingly apparent as the length of data records has increased [20]. In the US, studies of ecosystems started in the early 1900 s when forest, watershed, and rangeland sites were established, primarily by the United States Department of Agriculture (USDA) [21,22]. The research was both observation-based and experimental using manipulations related to management, such as altered fire frequency. Many long-term ecological sites now exist, including those in the Long Term Ecological

Corresponding author: Peters, D.P.C. (debpeter@nmsu.edu).

 0169-5347/$ – see front matter. Published by Elsevier Ltd. doi:10.1016/j.tree.2010.07.005 Trends in Ecology and Evolution 25 (2010) 592–601

Research Program (LTER) that began in 1980 [23] and sites studied by individuals or groups [24,25].

The 'ecology of the long' [15] complements detailed, process-based studies conducted over short time periods within a single ecosystem type (see next section: Ecology of the 'deep'). Ecological systems vary through time as environmental conditions change. Long-term data are needed to assess the rate and direction of change, to distinguish directional trends from short-term variability, and to determine effects of infrequent, yet extreme events and time lags in response [26–30]. Long-term data can inform government policy. For example, data showed an increase in acid rain in North America in the 1970 s [31], and that acid rain had negative impacts on forest growth and surface water chemistry [32,33]. These results led to the 1990 Amendments to the Clean Air Act which reduced sulfur dioxide emissions and sulfate concentrations in precipitation [34,35].

Comparisons of trends in drivers with ecological responses can infer causal relationships. For example, long-term studies off the Antarctic coast show strong correlations among drivers and system dynamics, and a state change from dominance by Adélie penguins to Gentoo and Chinstrap penguins (Figure 1). Now, short-term, detailed studies of predator–prey relationships under variable conditions of sea ice are needed to determine where, when, and how phytoplankton biomass or sea ice (or their interactions) drive loss of Adélie penguins or if a different set of processes are shifting dominance between penguin species, although field experiments at the required scale are challenging to conduct in this system. These short-term studies will need to be effectively integrated with the existing long-term studies if a complete understanding is to be achieved.

Limitations

Long-term observations can lead to hypotheses about processes underlying patterns, but cannot identify the processes. More than one process can create the same pattern, multiple interacting processes can result in the pattern, and spurious relationships can result with no causative explanation between pattern and process. In addition, the relationship between patterns and the processes driving them can change with temporal or spatial scale [41]. Long-term studies create challenges to data accessibility in that the sampling frequency and intensity, and the spatial scale (e.g. plot size) can change through time with turnover in personnel and as funding levels vary. Methods can change

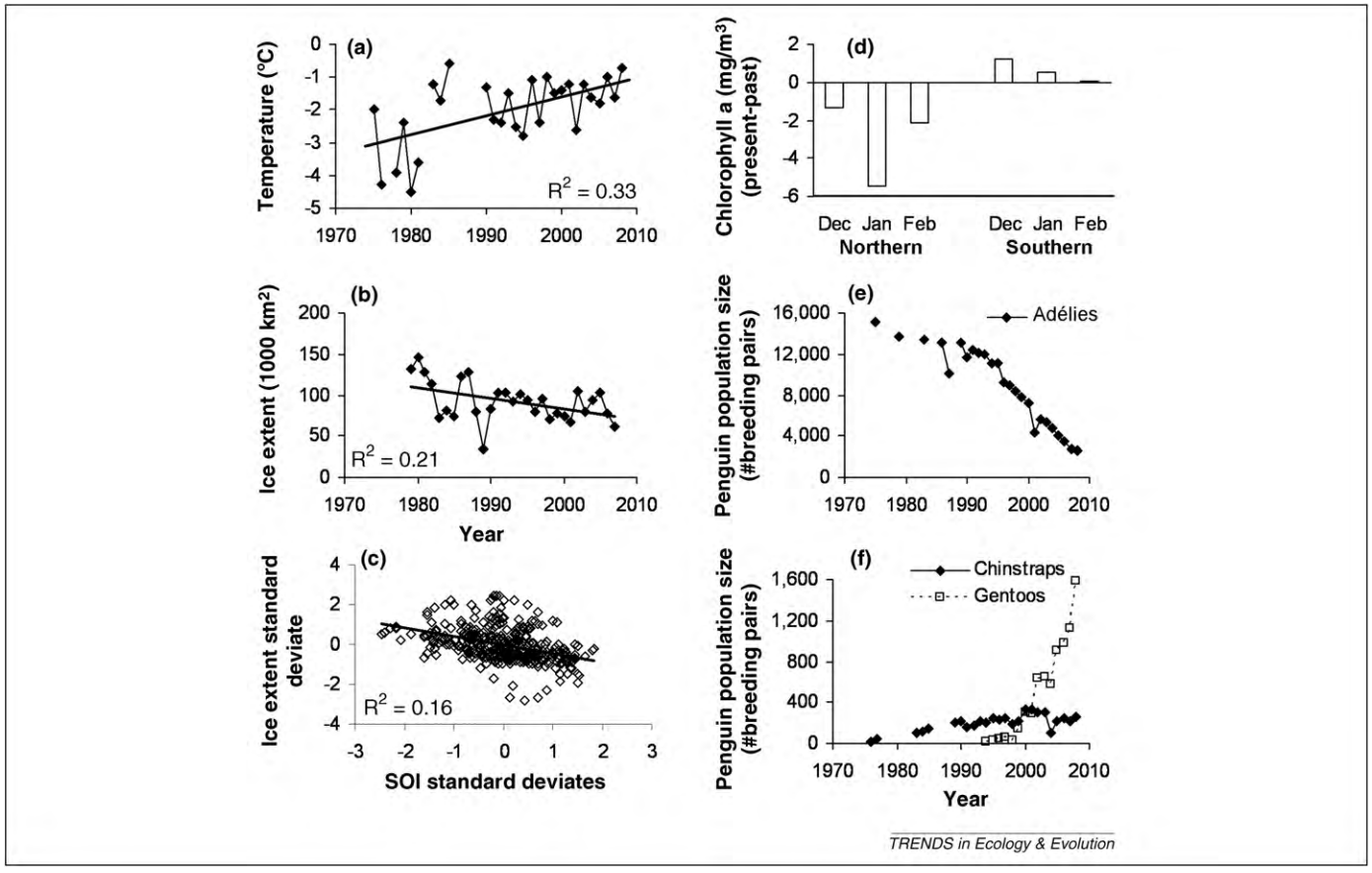


Figure 1. Long-term data for multiple drivers and ecological responses off the coast of the Western Antarctic Peninsula: **(a)** surface air temperatures have increased at some of the fastest recorded rates (temperature[°C] = –119 + 0.06 × Time[years]; $R^2 = 0.33$; p = 0.001) globally [36] (data from http://www.ecotrends.info). **(b)** Sea ice spatial extent has decreased significantly (ice extent[1000 km^2] = 2707 – 1.3 × Time[years]; $R^2 = 0.21$; p = 0.01) with a later advance and an earlier retreat of ice [37]. **(c)** Sea ice is related to the Southern Oscillation Index (SOI), and tends to advance during cooler La Nina periods, and retreat during warmer El Niño periods [38] (data from http://pal.lternet.edu shown as deviations from the mean: ice extent = –0.04 – 0.4 × SOI; $R^2 = 0.16$; $p < 0.0001$). **(d)** Phytoplankton biomass has shifted southward through time with decreases in the north (past: 1978–1986; present: 1998–2006) [39]. This shift in phytoplankton biomass is expected to reduce biomass of krill in the north, an important food source for Adélie penguins, **(e)** whose populations have been decreasing through time compared with **(f)** increases in populations of the ice-avoiding Gentoo and Chinstrap [40] (data from http://www.ecotrends.info). These patterns in drivers and biotic responses can be used to infer causal relationships, but identifying the key processes driving the state change between penguin species requires detailed studies of predator–prey relationships under multiple environmental conditions.

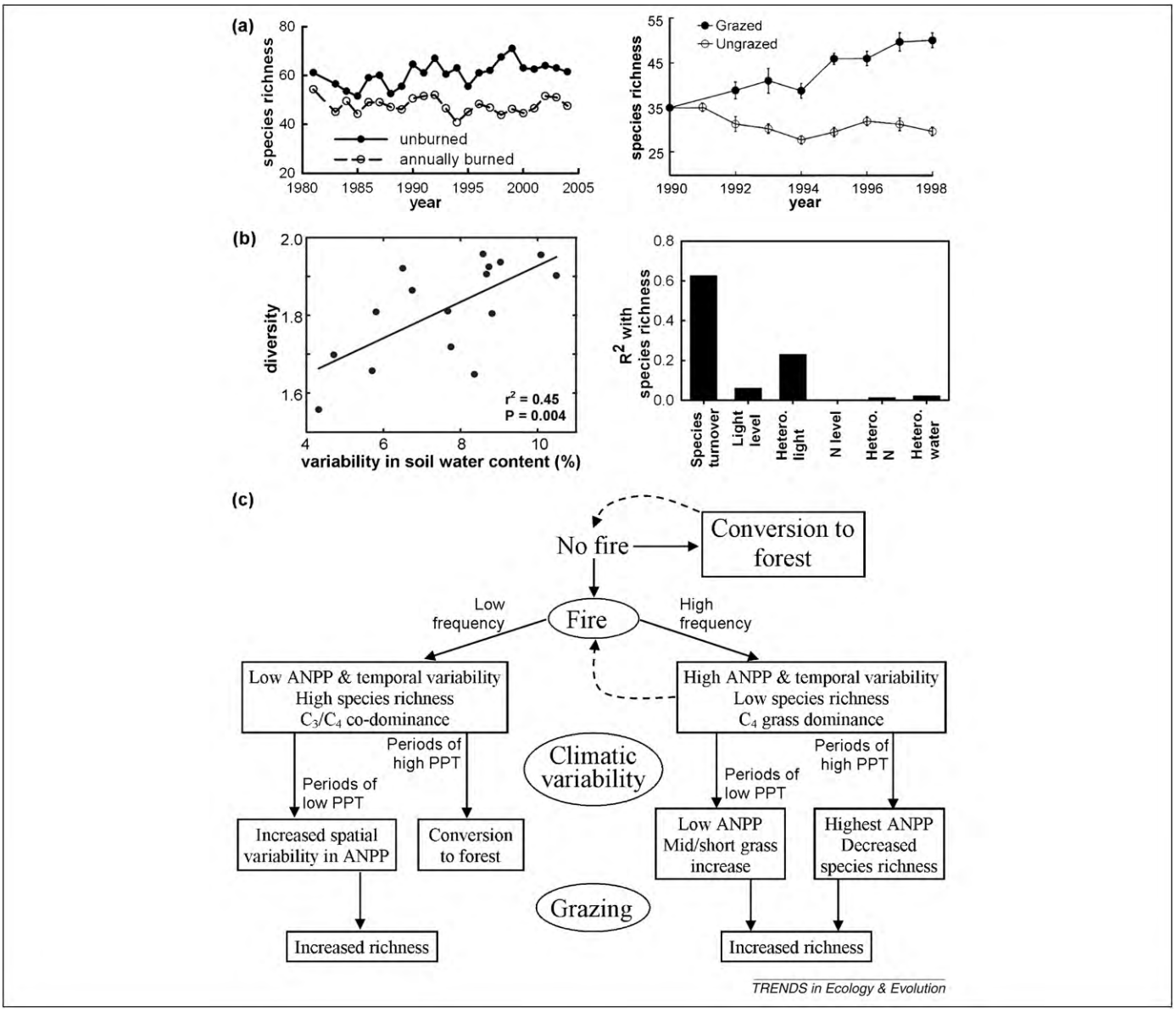


Figure 2. Deep understanding of tallgrass prairie at the Konza Prairie LTER site involves a suite of approaches. Short-term experiments are used to provide the mechanistic understanding for long-term observations, and a conceptual model is used to integrate the information. **(a)** Initial experiments focused on fire and grazing as historic drivers [51] and showed that [left] annual fire reduces plant species richness (updated from [52]) whereas [right] large herbivores (bison) increase plant and animal (not shown) species richness through time [53,54] (data from http://www.ecotrends.info). **(b)** [Left panel]: the hypothesis that variability in richness (and aboveground net primary production [ANPP], not shown) is related to variability in precipitation was tested using a short-term study where fewer, large rain events compared to natural rain events were added each year for four years [55]. This within- and between-year variability in rainfall increased variability in soil water with positive effects on plant diversity, a measure of species richness [55; Reprinted with permission from AAAS]. [Right panel]: short-term studies also showed that grazing increases spatial heterogeneity in light available to plants and increases turnover rate of species through time to result in higher species richness (redrawn from [56]) regardless of fire frequency [57]. **(c)** Results from these short- and long-term studies and others led to a conceptual framework used to: integrate information, test hypotheses, predict future dynamics, and strategically guide future research [redrawn and simplified from 50]. Extrapolation of these results to other sites in the tallgrass prairie requires information on spatial and temporal variability in both drivers and ecological responses.

with technological advances (e.g. automated sensors). Legacy data may not be well documented or in digital format, and variable names and file formats can change through time [9].

Ecology of the 'deep'

Place-based research conducted at one site or within one ecosystem type can provide deep understanding of processes underlying observed patterns [33]. Most studies are short-term (< 4 years), and some are conducted within a long-term context. These studies can also evaluate pattern–process relationships within and across scales [42–45]. Deep understanding is the hallmark of sites in the LTER Program where researchers test alternative hypotheses about drivers and responses using short-term experiments that lead logically from long-term observations, e.g. [46–49], and provide the mechanistic understanding for these observations.

For example, research at the Konza Prairie LTER site in Kansas has focused on teasing apart the relative importance of three drivers (fire, grazing, and climatic variability) in the dynamics of tallgrass prairie [50]. Short-term studies are used to examine the key mechanisms underlying long-term trends in observations. The LTER project began by observing grassland responses through time in response to manipulated fire frequency and grazing inten-

sity by large native herbivores (bison) under natural climatic variability (Figure 2a). High temporal variability in plant species richness (Figure 2a [left]) was hypothesized to reflect variation in soil water as affected by precipitation. To test this hypothesis, within season rainfall variability (fewer, larger rain events) was manipulated with no increase in rainfall amount. Soil water dynamics increased in variability, both within and among years, with positive effects on richness (Figure 2b [left]). The increase in plant (Figure 2a [right]) and consumer richness (not shown) under grazing was explained using short-term sampling that showed greater spatial heterogeneity in light available to plants and greater turnover of species in grazed than ungrazed areas (Figure 2b [right]). As a result of these and many other studies (http://www.konza.ksu.edu), the Konza Prairie LTER program developed a conceptual framework that integrates the effects of fire, grazing, and climatic variability on dynamics of these grasslands (Figure 2c).

Limitations

Site-based studies conducted without long-term observations can have limited generality because the temporal and spatial contexts of the results are unknown. Extrapolation of results from one site to another or to the region as a whole requires information on spatial and temporal variability in drivers and responses [58]. Site-based studies are insufficient to understand how ecosystems are connected by interactions among air, water, and land at broad scales [58,59]. Accessibility of data can be challenging if standard protocols of collection, archival, and retrieval are not followed [60].

Ecology of the 'broad'

Observation networks of sites collecting similar data across broad spatial extents have been operational in the US since at least 1830 with the census of human populations (http://www/census.gov). The National Weather Service started collecting meteorological data in 1870 (http://www.nws.noaa.gov/), and streamflow has been monitored at some sites for over 100 years (http://waterdata.usgs.gov). Observation networks have emerged over the past decade to collect ecological data using standard protocols, including the Ocean Observatories Initiative (OOI) [61], WATERS (http://www.watersnet.org), and NEON [16]. Other networks are collections of sites with similar missions, such as the USDA Agricultural Research Service (ARS) rangeland sites and the Forest Service (FS) experi-

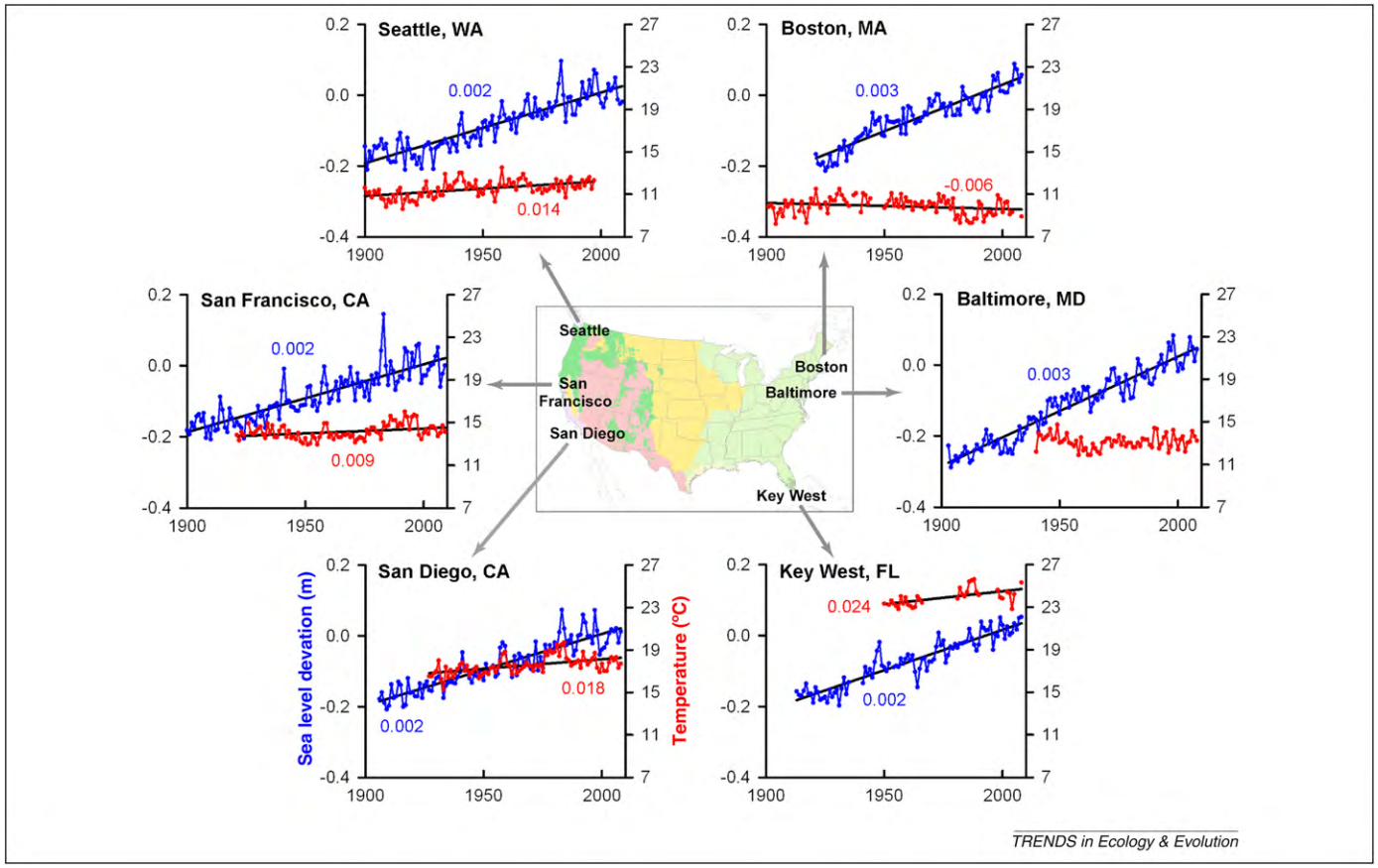


Figure 3. Broad-scale patterns can be observed using networks of sites either coordinated to collect similar data with standard protocols or integrated via the post-collection standardization of similar data. Sea level measured by the US Geological Survey (http://tidesandcurrents.noaa.gov/) using standard methods and instruments were used to calculate trends through time for cities along the east and west coasts of the country [19]. Long-term climate data obtained from a different source (https://www.ncdc.noaa.gov/) were used to calculate trends in average air temperature for the same cities or nearby research sites. Significant regression lines and slopes ($p \leq 0.05$) are shown in blue (sea level) and red (air temperature). All panels share the same y-axes labels of sea level and temperature. Comparing trends in the two drivers shows that most coastal sites have experienced an increase in sea level of 2–3 mm/y over the past 100 years. All west coast sites and Key West, FL have also experienced increasing air temperatures at rates of 0.01–0.02 °C /y (condensed data from http://www.ecotrends.info). Understanding the processes driving these patterns through time and predicting ecological responses requires detailed studies of mechanisms, both at individual sites and across environmental gradients, to capture variation in drivers and the biota.

mental forests, that collect data with site-specific methods; standardization is required before comparisons can be made [19,22].

In some cases, individuals collect data which, when combined, cover broad areas. The Global Population Dynamics Database contains animal and plant population data collected by individuals [62]. The National Phenology Network (NPN) contains data collected by citizen scientists using standard protocols [63]. Observing networks can be defined by regions or ecosystem types where individual projects with different protocols are integrated, such as the Global Lakes Ecological Observatory Network [64].

An integration of datasets from multiple networks is needed to compare continental-scale variation in multiple drivers, and to identify regions where multiple drivers are interacting to affect human and natural systems [19]. For example, sea level is increasing along the east and west coasts of the US, and surface air temperature is also increasing for sites on the west coast (Figure 3). Interactions between drivers may result in unexpected impacts on human populations and ecosystems located along the land–ocean margin [4,65].

Limitations

For observational networks that collect data with similar instruments at each site, standardized, aggregated data are accessible through a common web site. Comparisons of different kinds of data across networks require knowledge of and access to multiple web sites, and manual integration and analysis. These observing systems have limited ability to forecast dynamics without a long-term record of change for historical context, and a deep mechanistic understanding of pattern–process relationships across scales.

Linking the long, deep, and broad: a synthesis framework for understanding and prediction

Synthesis involves the integration of disparate data with existing concepts and theories to yield new knowledge, insights, and explanations [66]. Synthesis creates emergent knowledge through novel combinations of information [8]. A framework for synthesis is being developed where general patterns and underlying mechanisms are emerging from finding, blending, and integrating large volumes of data collected as part of the three strategies discussed above (Box 1). The framework is being developed to make complex data collected from different sources, locations, and disciplines easily accessible to and understood by a broad audience, and to develop new approaches and solutions to global change problems. This framework has points of contact with recent synthesis frameworks, and combines their key conceptual elements [10] with software tool development and training [67]. However, the focus on improving data and knowledge accessibility to a broad and diverse user community, with applications to policy, management, and personal actions, distinguishes this framework from others. This framework has five steps that address key limitations in the above three strategies.

First, data collected from different sources (individuals, sites, and networks) need to be assembled into digital formats where they are available to others (Box 1). These data can be from short-term experiments, long-term studies, and broad-scale observations guided by conceptual frameworks (Figures 1–3). This step may involve a number of technologies to: convert manually collected data into digital format, download data from sensors, verify data for accuracy (either manually or through automated value-checking routines), enter data into a user-specified database, release data to standard repositories, and post data onto the internet.

Second, these source data need to be standardized to allow their integration into a common database, either virtually with internet links or physically into a single database (Box 1). Although standard methods of data collection and analysis have been developed [68], and standard variable names and protocols are used by some research programs and networks [69], integrating data from different sources and disciplines remains a challenge that often requires post-collection standardization [9]. Even variables with well-defined standards, such as air temperature, can be collected in a variety of ways (e.g. at different heights) with different temporal resolutions (e.g. hourly, minimum and maximum daily). In some cases, such as observational networks, the data are already in a standard format, but only for one type of driver (e.g. climate variables); these data need to be integrated in a coherent way with ecological response data. In other cases, such as biotic data (plants, animals, and soils), the data are in a variety of formats that need to be standardized before integration [60]. Publishing details of methods, complete datasets, and metadata as extensions to scientific papers (e.g. Ecological Archives: http://esapubs.org/archive), and adhering to community-based standards [9] can provide information necessary for new analyses. As part of the standardization process, much can be learned about the structure of diverse datasets that can provide feedbacks to the data collection and assembly process of Step 1.

Third, the source data need to be condensed into simplified formats using aggregations in time and space to result in derived data products (Box 1). Source data from complex experimental designs are collected at very short time steps (days, months, and seasons) or small spatial extents (square meters and hectares) that make comparisons difficult. For example, rainfall data collected manually on an event basis or continuously through instrumentation need to be summed as monthly or annual totals. Biotic data often have a complex format needed to capture spatial and temporal variation in an ecological system [70]. For example, aboveground net primary production (ANPP) for a grassland site is often estimated by collecting biomass data seasonally by species in quadrats. ANPP is estimated by subtracting initial biomass by species (averaged across quadrats) from final average biomass. Summing ANPP by species results in site ANPP that can be compared across sites [19]. Additional derived data-products are needed to easily view patterns in the aggregated data, including X-Y graphs, maps, animations, and statistical results. The aggregation and analysis process will need new software tools and quantitative analyses, and training of scientists and information specialists to use and develop these tools.

Box 1. A framework for science-driven synthesis

A developing framework for synthesis includes five major steps and four classes of products that will result in making complicated data easily accessible for understanding and prediction by a broad audience (Figure I).

First, three strategies of ecological research (long-term studies; short-term, process-based studies; and broad-scale observations) result in large amounts of source data collected by individuals, sites, and networks of sites in a variety of formats, units, temporal and spatial resolutions, and degrees of complexity that need to be assembled. These data are often variable in their quality in terms of the degree to which they have been checked and corrected for errors.

Second, these diverse datasets need to undergo quality assurance and control, and to be standardized and integrated into one database, either a virtual database with internet links or a physical database. Much will be learned about the structure of diverse datasets that will provide important feedbacks to the data collection process.

Third, these data need to be converted into common aggregations to simplify their temporal and spatial resolutions that will allow comparison across sites and studies, and to promote synthesis. Derived data products need to be created, including X–Y graphs, maps, animations, and statistical results. The aggregation and analysis process will require new software tools and quantitative analyses, and training of scientists and information specialists to use and develop these tools.

Fourth, these derived data products need to be combined with other knowledge sources, new technologies, and approaches to promote new interpretations and synthesis of the data. A broad user community will be needed that includes individuals (e.g. scientists, land managers, citizen scientists, and information managers), networks of sites (e.g. LTER, USDA, and NEON), synthesis centers (e.g. National Center for Ecological Analysis and Synthesis [NCEAS, http://www.nceas.ucsb.edu/], National Evolutionary Synthesis Center [NESCent, http://www.nescent.org/], National Institute for Mathematical and Biological Synthesis [NIMBioS, http://www.nimbios.org/], and Powell Center; http://powellcenter.usgs.gov/), and state and federal agencies working together. These activities need to provide important feedbacks to the collection of additional data as well as to the development of tools and expertise for future analyses.

Fifth, these interpretations will need to inform policies, practices, and actions, and provide feedbacks to the collection of additional data. New technologies will need to be developed, and training of scientists and information managers in synthetic research will be needed to meet the challenges associated with synthesis.

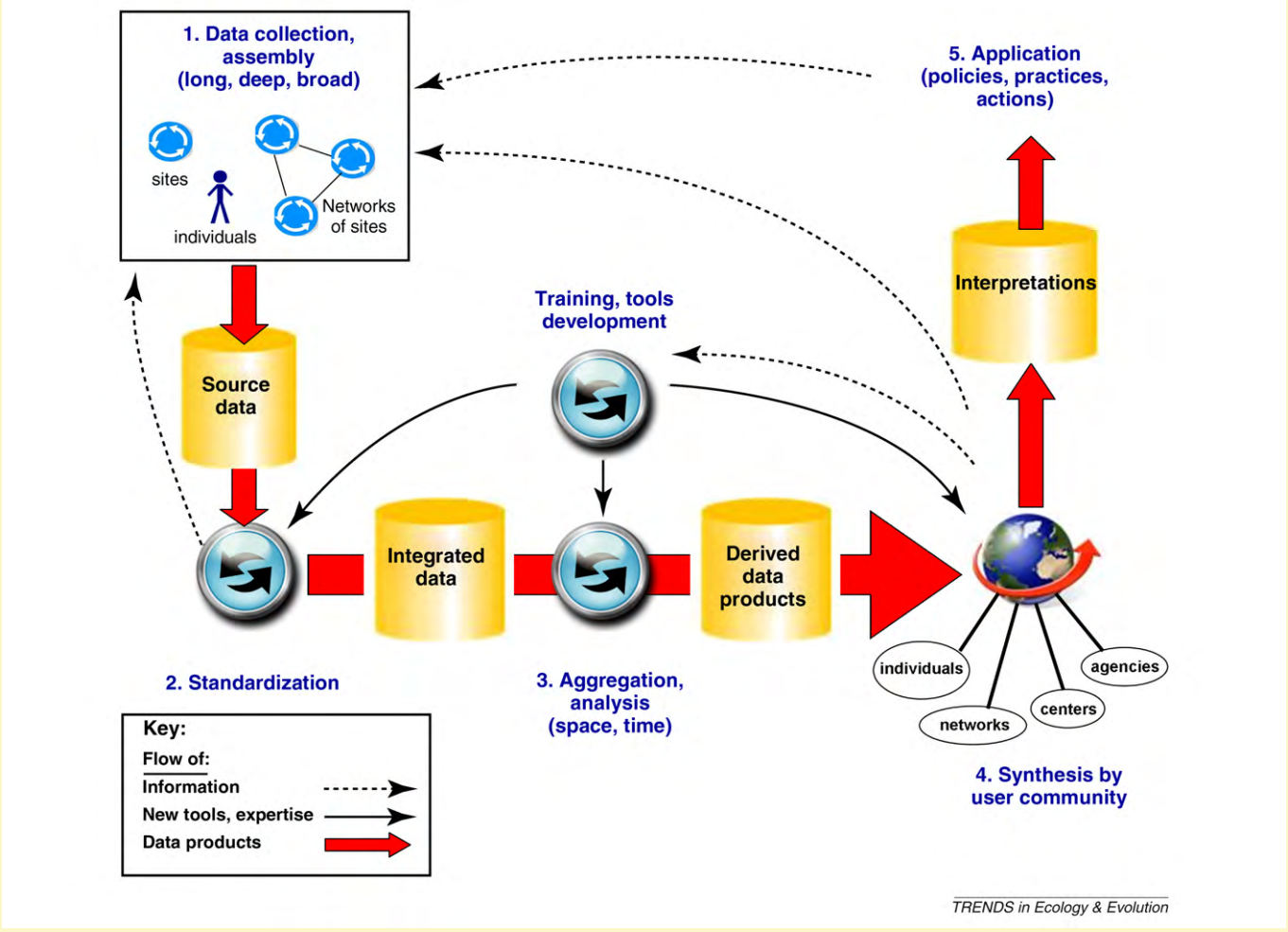


Figure I. A developing framework for synthesis includes five major steps (blue text) and four classes of products (yellow cylinders).

Although Steps 1–3 have been conducted on existing data in *post hoc* comparative analyses, e.g. [14,17,18], it is the fourth and fifth steps in this framework that have the potential to move synthesis to new quantitative levels required of current and future environmental problems.

Fourth, these derived data products need to be blended with other knowledge sources, new technologies, and approaches to promote new interpretations and syntheses by a broad user community (Box 1). Both traditional sources of knowledge (scientists, networks,

Box 2. EcoTrends as a first step towards a synthesis framework

Data from the EcoTrends Project (http://www.ecotrends.info) illustrate the first four steps of a developing framework for synthesis (Figure I). This project focuses on converting large volumes of *long-term data* from disparate sources into forms useful to others. Here, the data are used to address the following questions. How does continental- and regional-scale variation compare for trends in multiple drivers? What is the explanation for regional variation in drivers? What are the potential consequences of future changes in these drivers?

First, long-term data from three major sources were assembled: (1) broad-scale observation networks (National Climate Data Center [NCDC, http://www.ncdc.noaa.gov/], National Oceanic and Atmospheric Administration [NOAA, http://www.noaa.gov/], US Census Bureau [http://www.census.gov/], and National Atmospheric Deposition Program [NADP, http://nadp.sws.uiuc.edu/]); (2) individual investigators; and (3) monitoring data from research sites, primarily LTER, and the USDA-FS and USDA-ARS.

Second, these diverse data were corrected for errors, and integrated into a common database using standard formats, units, and variable names.

Third, the standardized source data were converted into common aggregations to simplify the temporal and spatial resolutions of the data. These derived data were viewed as graphs to answer our first question: high spatial variation in trends across the continent does not reflect regional patterns [19,58]. Within Colorado, patterns are highly variable for two sites (Niwot Ridge alpine site in the Rocky Mountains [NWT] and the Shortgrass Steppe [SGS] semiarid grassland site in the eastern plains): temperature and population density are increasing at both sites, although at different rates. Precipitation is not changing (not shown), and nitrogen deposition is either increasing (NWT) or not changing (SGS) (data from http://www.ecotrends.info).

Fourth, multiple datasets were used to interpret the data. To answer our second question, we needed information about the spatial location of the sites relative to cities. Increases in nitrogen deposition at NWT are likely related to increases in human population density in the Denver area, and upslope conditions that bring rainfall and atmospheric nitrogen from Denver to the mountains [76]. By contrast, the SGS site is located east of cities with slower rates of population increase, resulting in no change in nitrogen deposition through time. To answer our third question required information about biotic sensitivity to nitrogen: alpine sites, such as NWT, may be more negatively affected by nitrogen deposition in the future, as a result of increasing deposition rates and their sensitivity to nitrogen inputs [77] compared to SGS grasslands that are insensitive to nitrogen inputs without additional water [78].

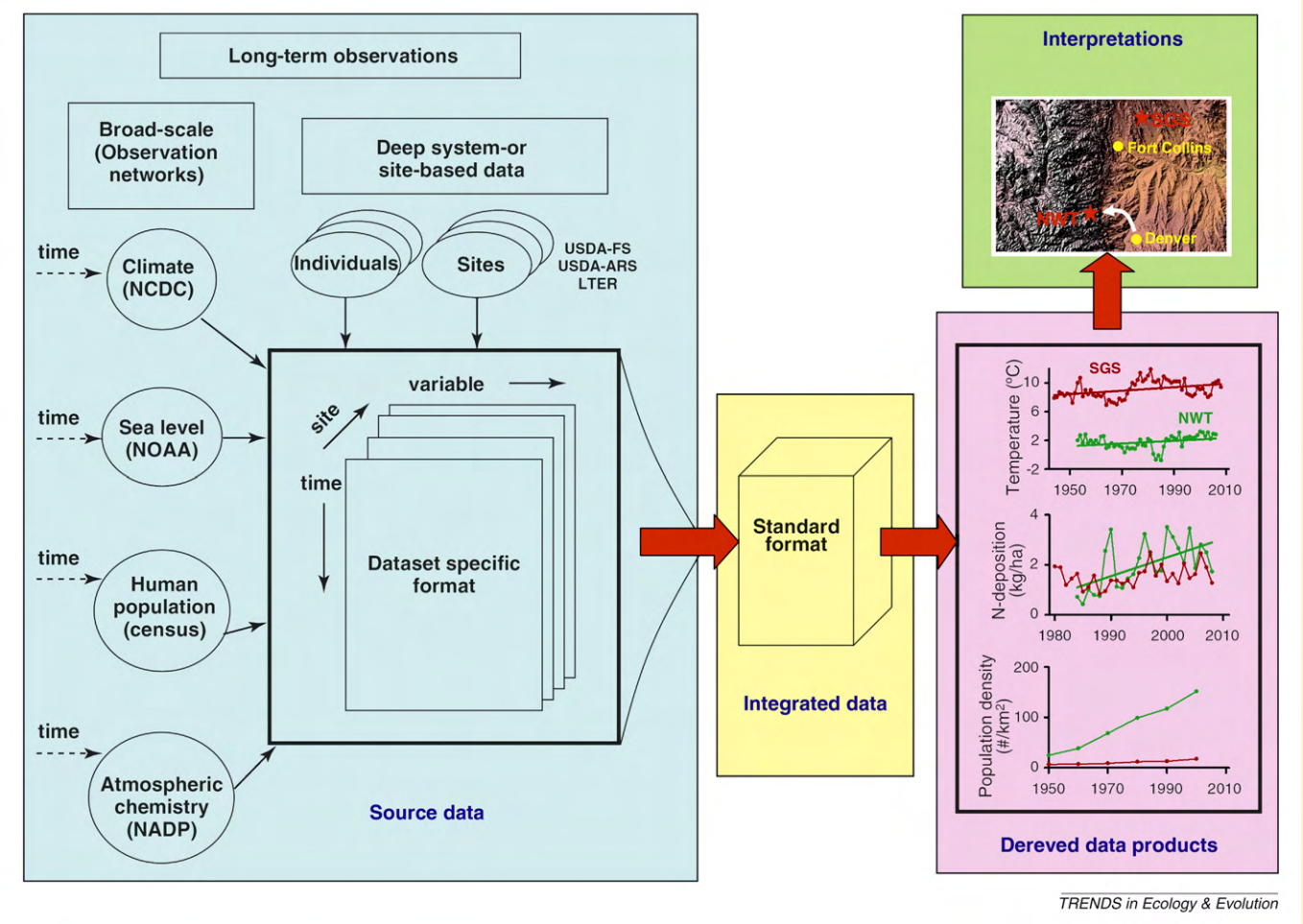


Figure I. Data from the EcoTrends Project (http://www.ecotrends.info) illustrate the first four steps of a developing framework for synthesis.

agencies) working as organized groups within structured (e.g. synthesis centers: NCEAS, NESCent, NIMBioS, and Powell Center) and unstructured environments (e.g. crowdsourcing [71]) are needed as well as other knowledge sources, such as citizen science initiatives (e.g. NPN; and North American Breeding Bird Survey, http://www.pwrc.usgs.gov/BBS/), and Traditional Ecological Knowledge [72]). A combination of quantitative and qualitative approaches and software development will be needed to blend diverse data, concepts, and theories from many disciplines. Training in using these new approaches will also be needed. These interpretations

need to provide important feedbacks to data collection and software development activities.

Fifth, these new interpretations need to inform policies, practices, and actions, and can be used directly to guide decision-making by individuals as well as by local, state, and federal policymakers (Box 1). In some cases, making data easily accessible and synthesized into best knowledge at the time may be insufficient to guide policy given other constraints (e.g. Kyoto Protocol and Copenhagen Accord: http://unfccc.int/). These applications, whether put into play or not, also need to provide feedbacks to data collection activities.

Applying the framework

The utility of this framework is illustrated by recent analyses from the EcoTrends Project (http://www.ecotrends.info). The aim of this project is to integrate and make easily accessible long-term data from many sources in four major categories: climate and climate-related drivers; air and stream water chemistry; human populations; and plants and animals [19]. At present, 50 US funded sites are included that represent ecosystems found globally (forests, grasslands, deserts, arctic, alpine, lakes, streams, coastal, urban). Here, key elements of the synthesis framework (Box 2) are used to show how *post hoc* comparisons of long-term data can be used to address the following scientific questions: How does continental- and regional-scale variation compare for trends in multiple drivers? What is the explanation for regional variation in drivers? What are the potential consequences of future changes in these drivers?

The continental US was selected as the broad-scale spatial unit, and the Rocky Mountains and eastern plains of Colorado were selected as the region for trends in four drivers: climate (precipitation and temperature), nitrogen deposition, and human population density. First, long-term source data for each driver were assembled from observation networks, research sites, and individuals. Second, data were tested and verified for completeness and accuracy, and then integrated into a standardized database. Third, derived data products were created by aggregating data into a common temporal unit (annual); the spatial unit was a site. Daily precipitation and event-based nitrogen deposition were summed for each year, and average daily temperature was averaged for each year. Human population density data on a decadal scale required no aggregation. The aggregated data were graphed through time for each site, and the trend based on the slope of a simple linear regression was calculated for each variable through time.

These comparable data were then used to answer our first question: high spatial variation in trends in air temperature and precipitation across the continent (not shown) does not necessarily reflect regional patterns [19,58]. Nitrogen emissions and deposition are higher on-average in the west compared to the east [19,73,74]. Human population density is increasing throughout the country, although rates over the past 50 years have been highest in the southwest and along the coasts [4,19,65]. Patterns are also highly variable within a region (Box 2). For example, temperature and population density are increasing, although at different rates for two sites in Colorado. Precipitation is not changing at either site (not shown), and nitrogen deposition is either increasing or not changing. These results can be used to guide decisions about air pollution mitigation for the increase in nitrogen in the mountains [75], and about global warming given the increase in temperature at both sites [2].

To answer our second question about explanations for this variability requires additional information in the fourth step. Specifically, information is needed about atmospheric sources of nitrogen and circulation patterns that affect nitrogen deposition (Box 2). Our third question about future consequences to ecological systems requires information, such as biotic sensitivity to nitrogen, and a synthesis of understanding about processes driving past patterns, how the drivers and ecosystems are changing, and how the past and present dynamics of ecosystems are likely to influence their future [79,80].

Prospects

Scientists have a responsibility to make their data accessible to others, where accessibility goes beyond making complex source data and metadata available on-line. The need for an understanding of scientific data by the public and decision-makers is critical if solutions to environmental problems are to find general acceptance [6,12,13]. A synthesis framework to integrate large volumes of complex data, often collected over long time periods, into coherent, easy-to-understand formats with other sources of knowledge shows great promise to link scientists with the rest of the world and to meet the challenges required by environmental problems. The framework allows general users to understand how drivers and responses are changing, and to critically examine the consequences of these changes and their personal actions to future dynamics of ecological systems.

Ecological knowledge obtained from traditional strategies (long-term studies; short-term process-based studies; and broad-scale observations) as well as non-traditional sources, such as citizen science initiatives and crowdsourcing, is invaluable to improved understanding and prediction through synthesis. These knowledge sources need to be integrated in novel, coherent ways to promote synthesis [10], and to strategically determine additional data needs [19]. More scientists need to be trained in quantitative synthesis, visualization and other software tools; assessments are limited more by there being few scientists trained in synthesis and communication than by deep knowledge of system dynamics [8,11]. The recent emergence of observation networks to capture variability across regions, continents, and oceans is important [16,61–64], but linking these networks with established sites and research programs for long-term context and deep understanding [42–49] is critical to optimizing resources with research needs.

Acknowledgements

This work was supported by National Science Foundation funding to New Mexico State University through the Jornada Basin LTER Program (DEB 06-18210) and to the University of New Mexico through the Sevilleta

LTER Program (DEB 06-20482) as well as funding by the USDA Agricultural Research Service to the Jornada Experimental Range. Individual and site contributors to the EcoTrends Project made their data and metadata available in a form useful for multi-site comparisons. The EcoTrends Project staff (Christine Laney and Jin Yao) assembled, integrated, and standardized the data into accessible formats and created figures. Members of the EcoTrends Editorial Committee provided valuable time and expertise to guide the development of derived data products. Kris Havstad, Scott Collins, Bill Fraser, and two anonymous reviewers provided helpful comments on the manuscript.

References

1 Millenium Ecosystem Assessment (2005) *Ecosystems and Human Well-being: Synthesis*, Island Press
2 Intergovernmental Panel on Climate Change Core Writing Team, Pachauri, R.K. and Reisinger, A. eds (2007) *Climate Change 2007: Synthesis Report*, Intergovernmental Panel on Climate Change
3 Parmesan, C. and Yohe, G. (2003) A globally coherent fingerprint of climate change impacts across natural systems. *Nature* 421, 37–42
4 Grimm, N.B. *et al.* (2008) The changing landscape: ecosystem responses to urbanization and pollution across climatic and societal gradients. *Front. Ecol. Environ.* 6, 264–272
5 Collins, J.P. and Crump, M.L. (2009) *Extinction in Our Times: Global Amphibian Decline*, Oxford University Press
6 Bennett, E.M. *et al.* (2005) Looking to the future of ecosystem services. *Ecosystems* 8, 125–132
7 The H. John Heinz Center (2008) *State of the Nation's Ecosystems*, Island Press
8 Carpenter, S.R. *et al.* (2009) Accelerate synthesis in ecology and environmental sciences. *BioScience* 59, 699–701
9 Laney, C.M. *et al.* (2011) Recommendations for data accessibility. In *Long-Term Trends in Ecological Systems: A Basis for Understanding Responses to Global Change* (Peters, D.P.C. *et al.* eds.), pp. 000-000, USDA Agricultural Research Service
10 Sidlauskas, B. *et al.* (2010) Linking big: the continuing promise of evolutionary synthesis. *Evolution* 64, 871–880
11 Frankel, F. and Reid, R. (2008) Distilling meaning from data. *Nature* 455, 30
12 NSF Advisory Committee for Environmental Research and Education. (2009) Transitions and Tipping points in Complex Environmental Systems: A Report by the NSF Advisory Committee Environmental Research and Education
13 Heinz Center (2008) *Environmental Information: A Road Map to the Future*, The H.J. Heinz III Center for Science, Economics, and the Environment
14 Knapp, A.K. *et al.* (2004) Generality in ecology: testing North American grassland rules in South African savannas. *Front. Ecol. Environ.* 2, 483–491
15 Carpenter, S.R. (2002) Ecological futures: building an ecology of the long now. *Ecology* 83, 2069–2083
16 Keller, M. *et al.* (2008) A continental strategy for the National Ecological Observatory Network. *Front. Ecol. Environ.* 6, 282–284
17 Abbott, I. and Le Maitre, D. (2010) Monitoring the impact of climate change on biodiversity: the challenge of megadiverse Mediterranean climate ecosystems. *Austral Ecol.* 35, 406–422
18 Morecroft, M.D. *et al.* (2009) The UK Environmental Change Network: emerging trends in the composition of plant and animal communities and the physical environment. *Biol. Conser.* 142, 2814–2832
19 Peters, D.P.C. *et al.*, eds (2011) *Long-Term Trends in Ecological Systems: A Basis for Understanding Responses to Global Change*, USDA Agricultural Research Service
20 Janzen, H.H. (2009) Long-term ecological sites: musings on the future, as seen (dimly) from the past. *Glob. Change Biol.* 15, 2770–2778
21 Lugo, A.E. *et al.* (2006) Long-term research at the USDA Forest Service's experimental forests and ranges. *BioScience* 56, 39–48
22 Moran, M.S. *et al.* (2008) Long-term data collection at USDA experimental sites for studies of ecohydrology. *Ecohydrology* 1, 377–393
23 Hobbie, J.E. *et al.* (2003) The US Long Term Ecological Research Program. *BioScience* 53, 21–32
24 Brown, J.H. *et al.* (2001) Complex species interactions and the dynamics of ecological systems: long-term experiments. *Science* 293, 643–650
25 McClaran, M.P. (2003) A century of vegetation change on the Santa Rita Experimental Range. In *Santa Rita Experimental Range: 100 Years (1903-2003) of Accomplishments and Contributions* (McClaran, M.P. et al. eds), pp. 16-33, RMRS-P-30, Rocky Mountain Research Station
26 Magnuson, J.J. (1990) Long-term ecological research and the invisible present. *BioScience* 40, 495–501
27 Kratz, T.K. *et al.* (2003) Ecological variability in space and time: insights gained from the US LTER Program. *BioScience* 53, 57–67
28 Covich, A.P. *et al.* (2006) Effects of drought and hurricane disturbance on headwater distributions of palaemonid river shrimp (*Macrobrachium* spp.) in the Luquillo Mountains, Puerto Rico. *J. North Amer. Benthol. Soc.* 25, 99–107
29 Lugo, A.E. (2008) Visible and invisible effects of hurricanes on forest ecosystems: an international review. *Austral Ecol.* 33, 368–398
30 Drew, A.P. *et al.* (2009) Sixty-two years of change in subtropical wet forest structure and composition at El Verde, Puerto Rico. *Interciencia* 34, 34–40
31 Likens, G.E. and Bormann, F.H. (1974) Acid rain: a serious environmental problem. *Science* 184, 1176–1179
32 Driscoll, C.T. *et al.* (2001) Acidic deposition in the northeastern United States: sources and inputs, ecosystem effects, and management strategies. *BioScience* 51, 180–198
33 Likens, G.E. (2004) Some perspectives on long-term biogeochemical research from the Hubbard Brook Ecosystem Study. *Ecology* 85, 2355–2362
34 Butler, T.J. *et al.* (2001) Regional-scale impacts of Phase I of the Clean Air Act Amendments in the USA: the relation between emissions and concentrations, both wet and dry. *Atmos. Environ.* 35, 1015–1028
35 Likens, G.E. *et al.* (2002) The biogeochemistry of sulfur at Hubbard Brook. *Biogeochemistry* 60, 235–316
36 Vaughn, D.G. *et al.* (2003) Recent rapid regional climate warming on the Antarctic Peninsula. *Clim. Change* 60, 243–274
37 Ducklow, H.W. *et al.* (2007) Marine ecosystems: The West Antarctic Peninsula. *Phil. Trans. Royal Soc. London B* 362, 67–94
38 Stammerjohn, S.E. *et al.* (2008) Trends in Antarctic annual sea ice retreat and advance and their relation to ENSO and Southern Annular Mode Variability. *J. Geophys. Res.* 113, C03S90, DOI: 10.1029/2007JC004269
39 Montes-Hugo, M. *et al.* (2009) Recent changes in phytoplankton communities associated with rapid regional climate change along the Western Antarctic Peninsula. *Science* 323, 1470–1473
40 McClintock, J. *et al.* (2008) Ecological impacts of climate change on the Antarctic Peninsula. *Am. Sci.* 96, 302–310
41 Levin, S.A. (1992) The problem of pattern and scale in ecology. *Ecology* 73, 1943–1967
42 Peters, D.P.C. *et al.* (2006) Disentangling complex landscapes: new insights to forecasting arid and semiarid system dynamics. *BioScience* 56, 491–501
43 Allen, C.D. (2007) Interactions across spatial scales among forest dieback, fire, and erosion in northern New Mexico landscapes. *Ecosystems* 10, 797–808
44 Willig, M.R. *et al.* (2007) Cross-scale responses of biodiversity to hurricane and anthropogenic disturbance in a tropical forest. *Ecosystems* 10, 824–838
45 Young, D.R. *et al.* (2007) Cross-scale patterns in shrub thicket dynamics in the Virginia barrier complex. *Ecosystems* 10, 854–863
46 Chapin, F.S. *et al.* (2005) *Alaska's Changing Boreal Forest*, Oxford University Press
47 Magnuson, J.J. *et al.* (2005) *Long-Term Dynamics of Lakes in the Landscape*, Oxford University Press
48 Havstad, K.M. *et al.*, eds (2006) *Structure and Function of a Chihuahuan Desert Ecosystem: The Jornada Basin LTER*, Oxford University Press
49 Lauenroth, W.K. and Burke, I.C., eds (2008) *Ecology of the Shortgrass Steppe*, Oxford University Press
50 Knapp, A.K. *et al.* (1998) *Grassland Dynamics*, Oxford University Press
51 Daubenmire, R. (1968) Ecology of fire in grasslands. *Adv. Ecol. Res.* 5, 209–266
52 Briggs, J.M. and Knapp, A.K. (1995) Interannual variability in primary production in tallgrass prairie: climate, soil moisture, topographic position and fire as determinants of aboveground biomass. *Am. J. Bot.* 82, 1024–1030
53 Towne, E.G. *et al.* (2005) Vegetation trends in tallgrass prairie from bison and cattle grazing. *Ecol. Appl.* 15, 1550–1559

54 Joern, A. (2005) Disturbance by fire frequency and bison grazing modulate grasshopper species assemblages (Orthoptera) in tallgrass prairie. *Ecology* 86, 861–873
55 Knapp, A.K. *et al.* (2002) Rainfall variability, carbon cycling, and plant species diversity in a mesic grassland. *Science* 298, 2202–2205
56 Bakker, C. *et al.* (2003) Does resource availability, resource heterogeneity or species turnover mediate changes in plant species richness in grazed grasslands? *Oecologia* 137, 385–391
57 Collins, S.L. *et al.* (1998) Modulation of diversity by grazing and mowing in native tallgrass prairie. *Science* 280, 745–747
58 Peters, D.P.C. *et al.* (2008) Living in an increasingly connected world: a framework for continental-scale environmental science. *Front. Ecol. Environ.* 6, 229–237
59 Adger, W.N. *et al.* (2009) Nested and teleconnected vulnerabilities to environmental change. *Front. Ecol. Environ.* 7, 150–157
60 Michener, W.K. and Brunt, J.W., eds (2000) *Ecological Data: Design, Management and Processing*, Blackwell Science Ltd
61 Clark, H.L. and Isern, A. (2003) The OOI and the IOOS – can they be differentiated? An NSF perspective. *Oceanography* 16, 20–21
62 Inchausti, P. and Halley, J. (2001) Investigating long-term ecological variability using the global population dynamics database. *Science* 293, 655–657
63 Betancourt, J.L. (2005) Implementing a U.S. National Phenology Network. *Eos* 86, 539–541
64 Hamilton, D.P. *et al.* (2006) *Development of a Global Lake Ecological Observatory Network*, Institute of Industrial Science, University of Tokyo, Japan and Lake Biwa Environmental Research Institute
65 Hopkinson, C.S. *et al.* (2008) Forecasting effects of sea-level rise and windstorms on coastal and inland ecosystems. *Front. Ecol. Environ.* 6, 255–263
66 Pickett, S.T.A. *et al.* (2007) *Ecological Understanding: the Nature of Theory and the Theory of Nature*, (2nd edn), Academic Press
67 Adelman, S.J. *et al.* (2004) Understanding environmental complexity through a distributed knowledge network. *BioScience* 54, 240–246
68 Robertson, G.P. *et al.* (1999) *Standard Soil Methods for Long-term Ecological Research*, Oxford University Press
69 Baker, K.S. *et al.* (2000) Evolution of a multisite network information system: the LTER information management paradigm. *BioScience* 50, 963–978
70 Fahey, T.J. and Knapp, A.K., eds (2007) *Principles and Standards for Measuring Primary Production*, Oxford University Press
71 Shirky, C. (2008) *Here Comes Everybody*, Penguin Press
72 Berkes, F. *et al.* (2000) Rediscovery of traditional ecological knowledge as adaptive management. *Ecol. Appl.* 10, 1251–1262
73 Driscoll, C.T. *et al.* (2011). Cross-site comparisons of precipitation and surface water chemistry. In *Long-Term Trends in Ecological Systems: A Basis for Understanding Responses to Global Change* (Peters, D.P.C. *et al.*, eds.), pp. 000-000, USDA Agricultural Research Service
74 Fenn, M.E. *et al.* (2003) Nitrogen emissions, deposition, and monitoring in the western United States. *BioScience* 53, 391–403
75 Williams, M.W. and Tonnessen, K.A. (2000) Critical loads for inorganic nitrogen deposition in the Colorado Front Range. *USA. Ecol. Appl.* 10, 1648–1665
76 Burns, D.A. (2003) Atmospheric nitrogen deposition in the Rocky Mountains of Colorado and southern Wyoming – a review and new analysis of past study results. *Atmos. Environ.* 37, 921–932
77 Baron, J.S. *et al.* (2005) High elevation ecosystem responses to atmospheric deposition of nitrogen in the Colorado Rocky Mountains, USA. In *Global Change and Mountain Regions* (Huber, U.M. *et al.*, eds), pp. 429–436, Springer
78 Lauenroth, W.K. *et al.* (1978) The effects of water and nitrogen induced stresses on plant community structure in a semiarid grassland. *Oecologia* 36, 211–222
79 Coreau, A. *et al.* (2009) The rise of research on futures in ecology: rebalancing scenarios and predictions. *Ecol. Lett.* 12, 1277–1286
80 Jackson, S.T. *et al.* (2009) Ecology and the ratchet of events: climate variability, niche dimensions, and species distributions. *Proc. Natl. Acad. Sci. U. S. A.* 106, 19685–19692