

ECOLOGICAL RESEARCH FROM A TRULY LONG-TERM PERSPECTIVE:  
WHAT NEEDS TO BE DONE TODAY TO SUPPORT ECOLOGICAL ANALYSIS IN 2100

# Data Management for LTER @ 100

---

Robert J. Robbins

*UCSD (sort of)*

RJR8222@gmail.com

ECOLOGICAL RESEARCH FROM A TRULY LONG-TERM PERSPECTIVE:  
WHAT NEEDS TO BE DONE TODAY TO SUPPORT ECOLOGICAL ANALYSIS IN 2100

## DATA MANAGEMENT FOR LTER: 1980 – 2010

A POSITION PAPER

*prepared by*

ROBERT J. ROBBINS

*in conjunction with the NSF thirty-year review of LTER*



<http://www.nsf.gov/pubs/2012/bio12002/bio12002.pdf>

### ECOLOGICAL RESEARCH FROM A TRULY LONG-TERM PERSPECTIVE: WHAT NEEDS TO BE DONE TODAY TO SUPPORT ECOLOGICAL ANALYSIS IN 2100

was prepared in conjunction with a twenty-year review.

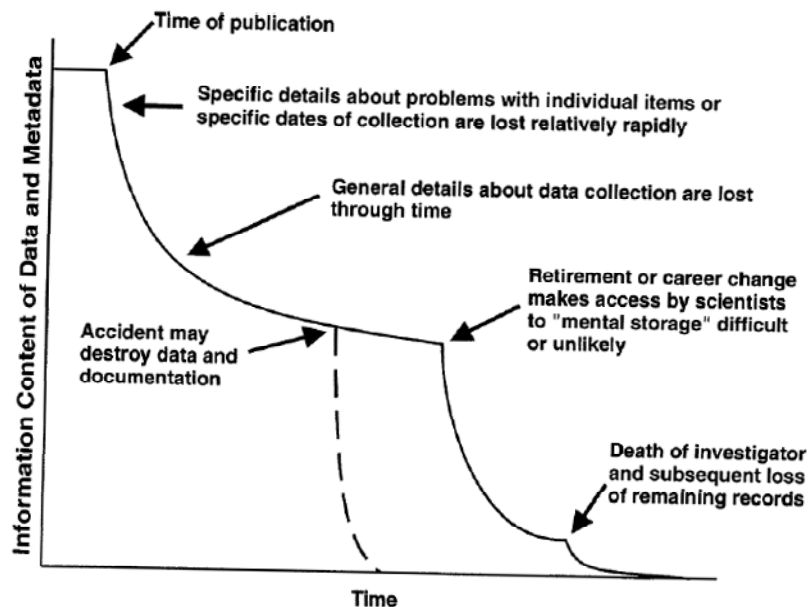
When the committee started its review, Joann Roskoski, the acting AD of the Directorate for Biological Sciences, urged us to think big:

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.

## ECOLOGICAL RESEARCH FROM A TRULY LONG-TERM PERSPECTIVE: WHAT NEEDS TO BE DONE TODAY TO SUPPORT ECOLOGICAL ANALYSIS IN 2100



**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)

### ECOLOGICAL RESEARCH FROM A TRULY LONG-TERM PERSPECTIVE: WHAT NEEDS TO BE DONE TODAY TO SUPPORT ECOLOGICAL ANALYSIS IN 2100

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.

Level	Planned use			
III	Publishable & auditable	Inadequate	Minimal	Good Practice
II	Searchable & third party reuse	Minimal	Good Practice	Excessive
I	Exchange with expert colleague	Good Practice	Excessive	Excessive
		Low	Medium	High
		Free format, ASCII, narrative, or hard copy	Mixed format, partially parameterized	Fixed format, highly parameterized, self-documenting and automatically parsable.
Amount of Structure (Level of effort)				

**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)

ECOLOGICAL RESEARCH FROM A TRULY LONG-TERM PERSPECTIVE:  
WHAT NEEDS TO BE DONE TODAY TO SUPPORT ECOLOGICAL ANALYSIS IN 2100

## Happy Customers: An Aside

**Most people classify all IT resources into one of two categories:**

## Happy Customers: An Aside

**Most people classify all IT resources into one of two categories:**

- **The stuff I personally use**
- **The stuff other people use**

## Happy Customers: An Aside

**Most people classify all IT resources into one of two categories:**

- **The stuff I personally use** (which, although absolutely mission critical, is out of date, woefully underpowered and must be upgraded and expanded immediately).
- **The stuff other people use**



## Happy Customers: An Aside

**Most people classify all IT resources into one of two categories:**

- **The stuff I personally use** (which, although absolutely mission critical, is out of date, woefully underpowered and must be upgraded and expanded immediately).
- **The stuff other people use** (which, although really an unnecessary luxury, is incredibly over-powered, way too expensive, and should be controlled or even eliminated immediately).

## Happy Customers: An Aside

**Most people classify all IT resources into one of two categories:**

- **The stuff I personally use** (which, although absolutely mission critical, is out of date, woefully underpowered and must be upgraded and expanded immediately).
- **The stuff other people use** (which, although really an unnecessary luxury, is incredibly over-powered, way too expensive, and should be controlled or even eliminated immediately).

### ECOLOGICAL RESEARCH FROM A TRULY LONG-TERM PERSPECTIVE: WHAT NEEDS TO BE DONE TODAY TO SUPPORT ECOLOGICAL ANALYSIS IN 2100

and increasing speed — e...ays more...are added to GenBank and  
added in the first ten years.<sup>2</sup>

*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

ECOLOGICAL RESEARCH FROM A TRULY LONG-TERM PERSPECTIVE:  
WHAT NEEDS TO BE DONE TODAY TO SUPPORT ECOLOGICAL ANALYSIS IN 2100

## **Funding Science vs funding infrastructure:**

- Where does the money come from?
- Who has a fiduciary (or ethical or moral) duty to do what?

## Funding Science:

- Program staff have a duty to taxpayers to deliver the best science possible for the money expended.
- Program staff have a duty to the community to presiding over an open and fair competition, levelling the playing field for all, ensuring an adequate and fair assessment, then funding the strongest proposals.

## Funding Infrastructure:

- Program staff have a duty to taxpayers to deliver the best scientific infrastructure possible for the money expended.
- Program staff have a duty to the community to procure, on behalf of the community, the best (and most appropriate) scientific infrastructure possible for the money expended.

---

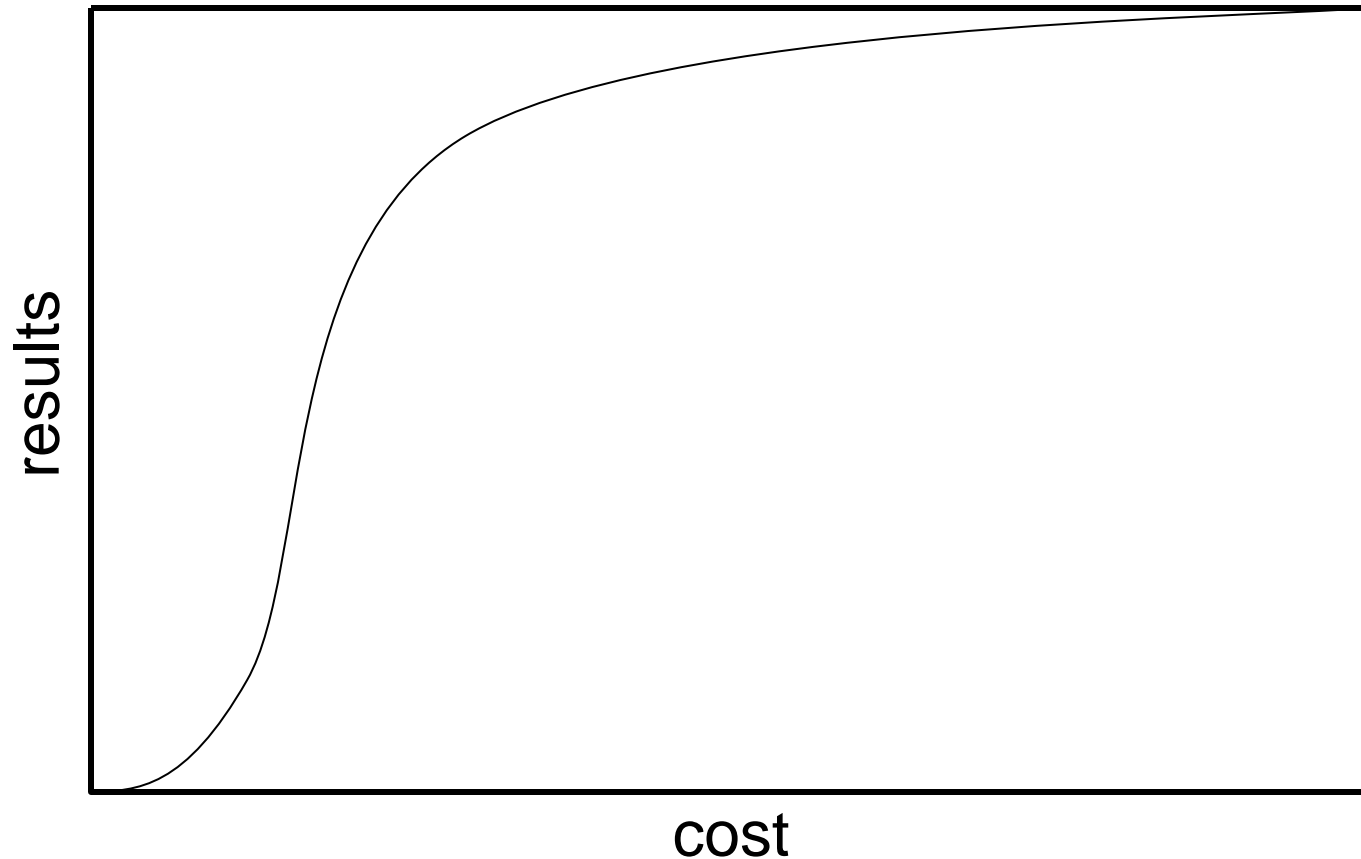
ECOLOGICAL RESEARCH FROM A TRULY LONG-TERM PERSPECTIVE:  
WHAT NEEDS TO BE DONE TODAY TO SUPPORT ECOLOGICAL ANALYSIS IN 2100

## LTER:

- The conduct of LTER research is science.
- The development of long-term data resources for others to use is infrastructure.

ECOLOGICAL RESEARCH FROM A TRULY LONG-TERM PERSPECTIVE:  
WHAT NEEDS TO BE DONE TODAY TO SUPPORT ECOLOGICAL ANALYSIS IN 2100

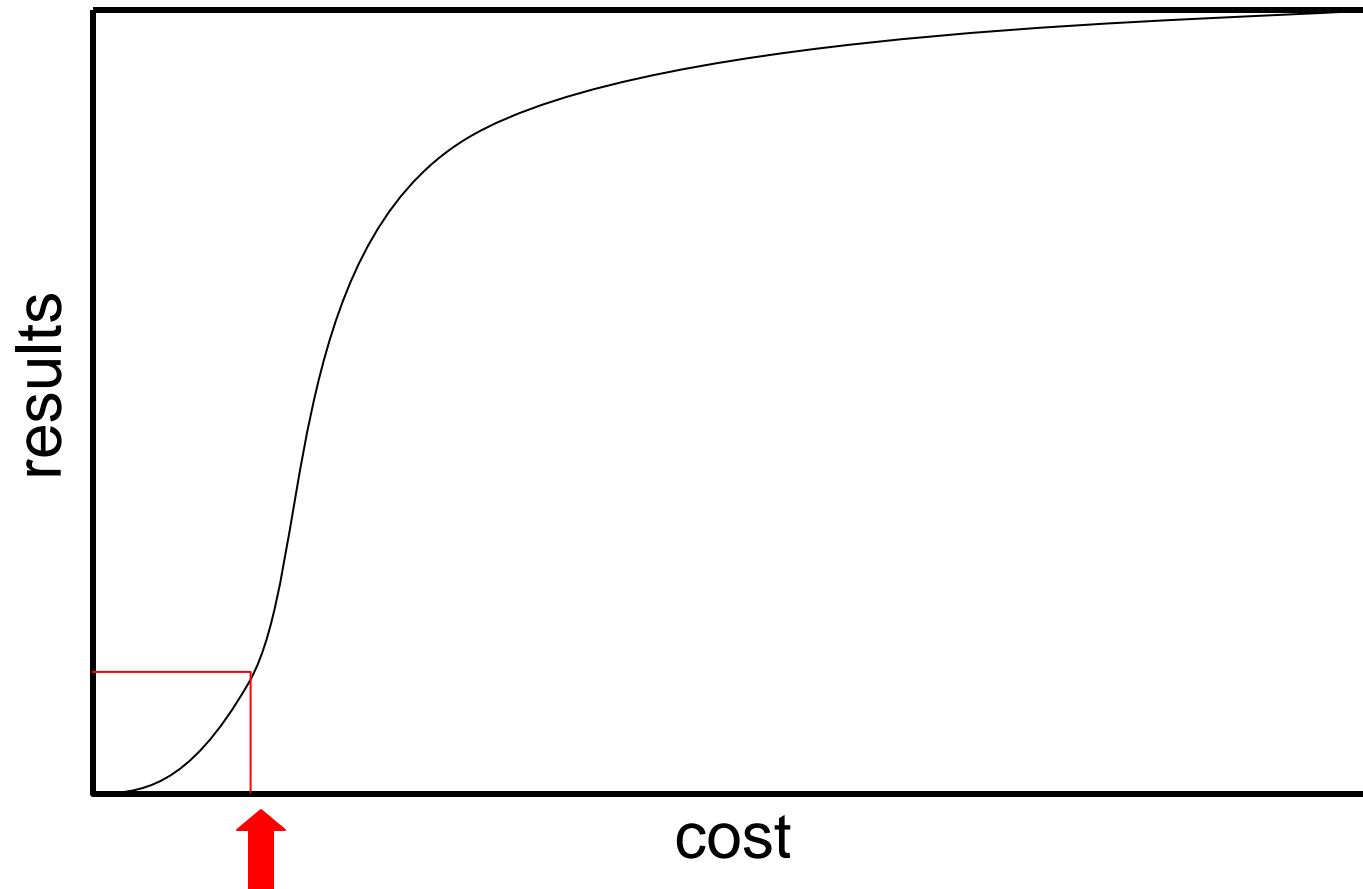
## The 80:20 Rule





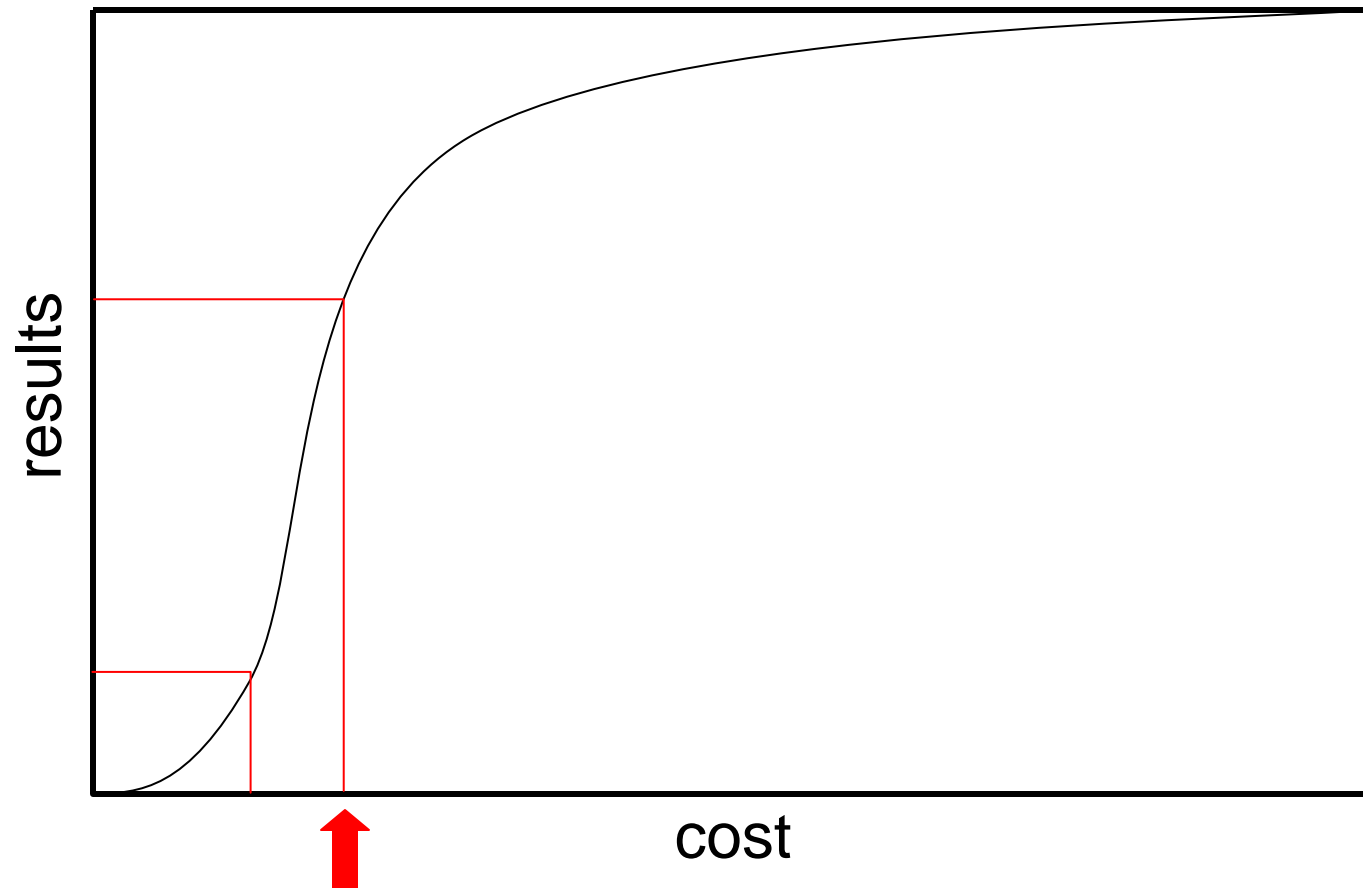
ECOLOGICAL RESEARCH FROM A TRULY LONG-TERM PERSPECTIVE:  
WHAT NEEDS TO BE DONE TODAY TO SUPPORT ECOLOGICAL ANALYSIS IN 2100

## The 80:20 Rule



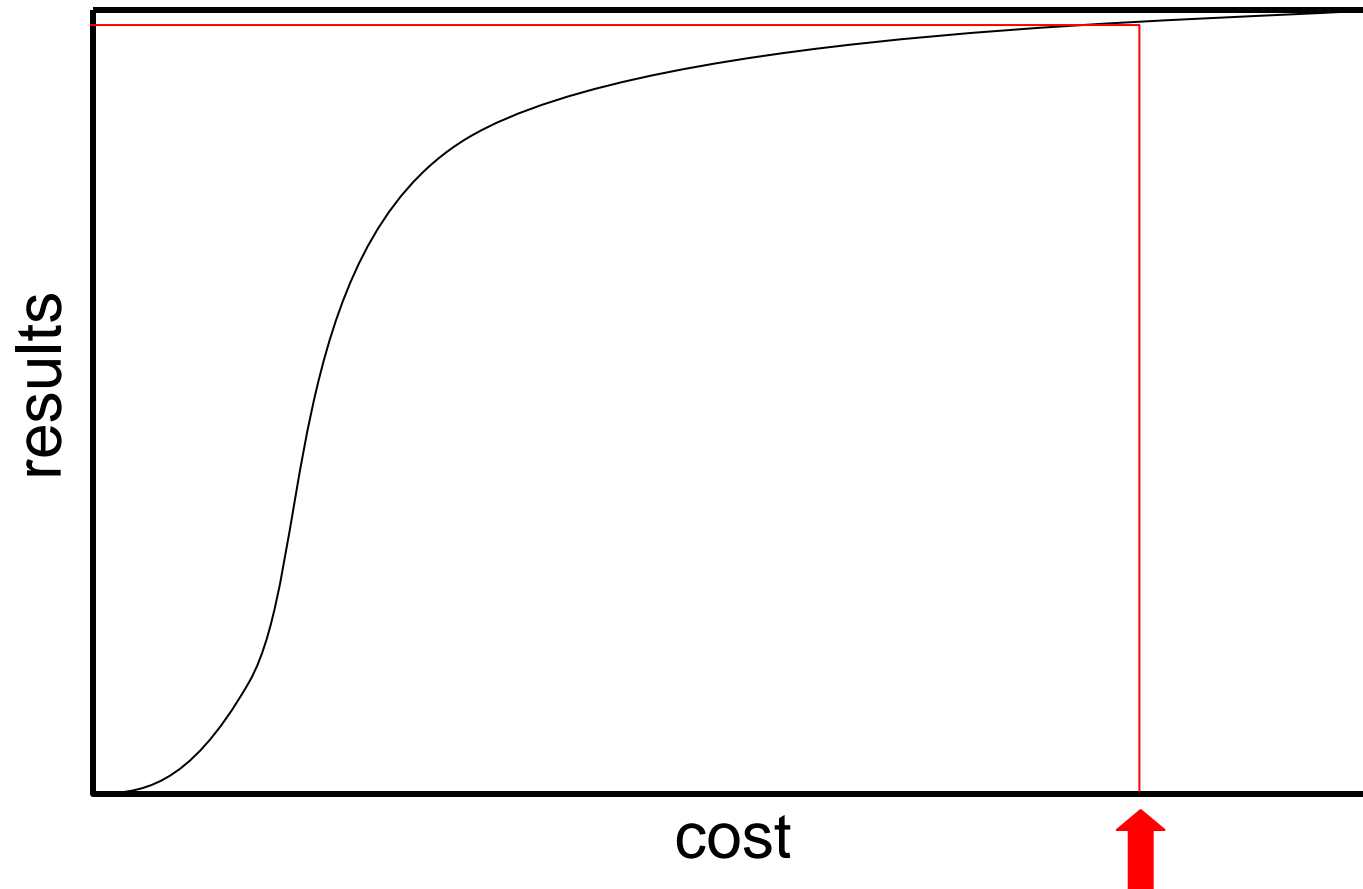
ECOLOGICAL RESEARCH FROM A TRULY LONG-TERM PERSPECTIVE:  
WHAT NEEDS TO BE DONE TODAY TO SUPPORT ECOLOGICAL ANALYSIS IN 2100

## The 80:20 Rule



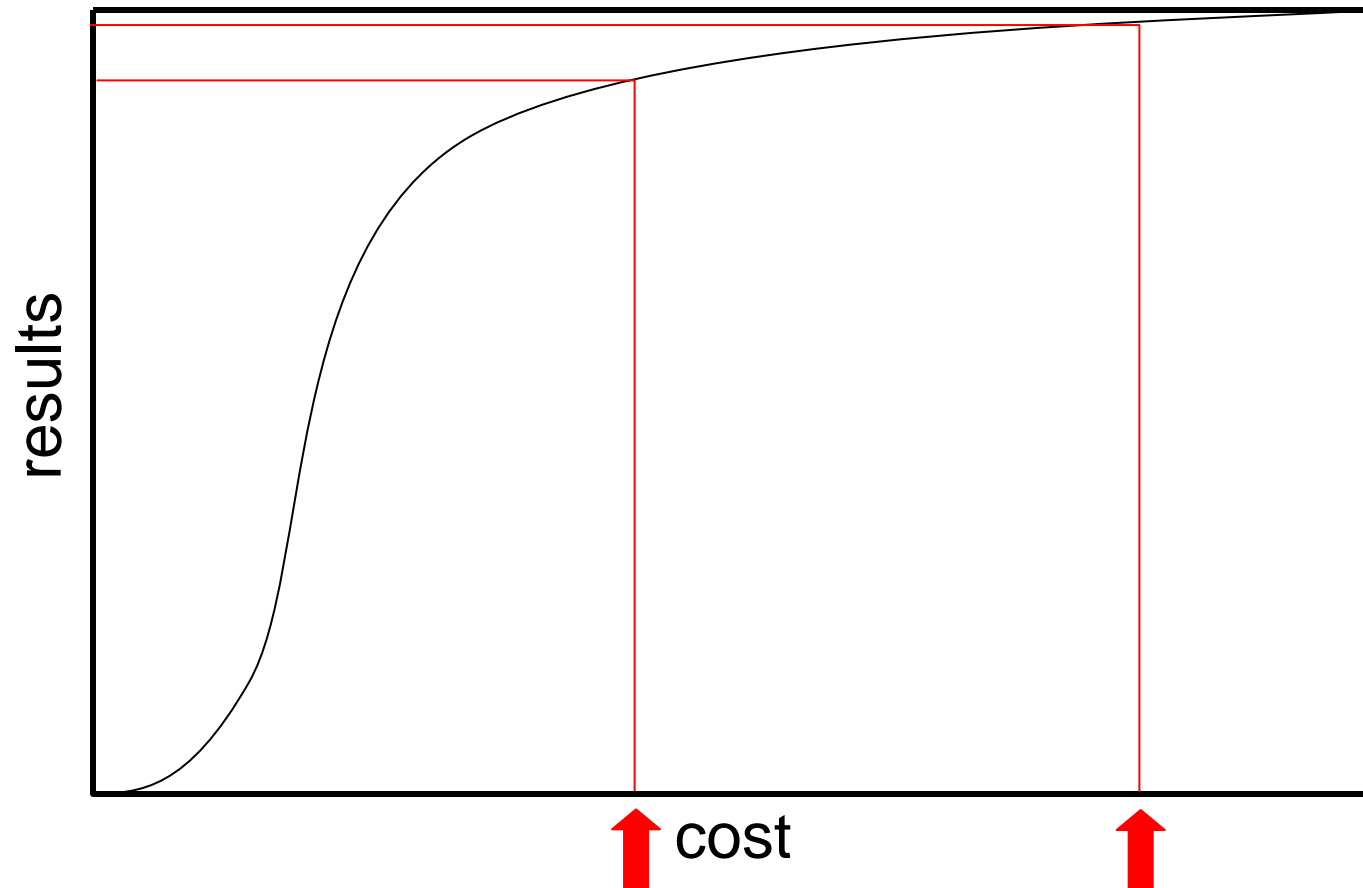
ECOLOGICAL RESEARCH FROM A TRULY LONG-TERM PERSPECTIVE:  
WHAT NEEDS TO BE DONE TODAY TO SUPPORT ECOLOGICAL ANALYSIS IN 2100

## The 80:20 Rule



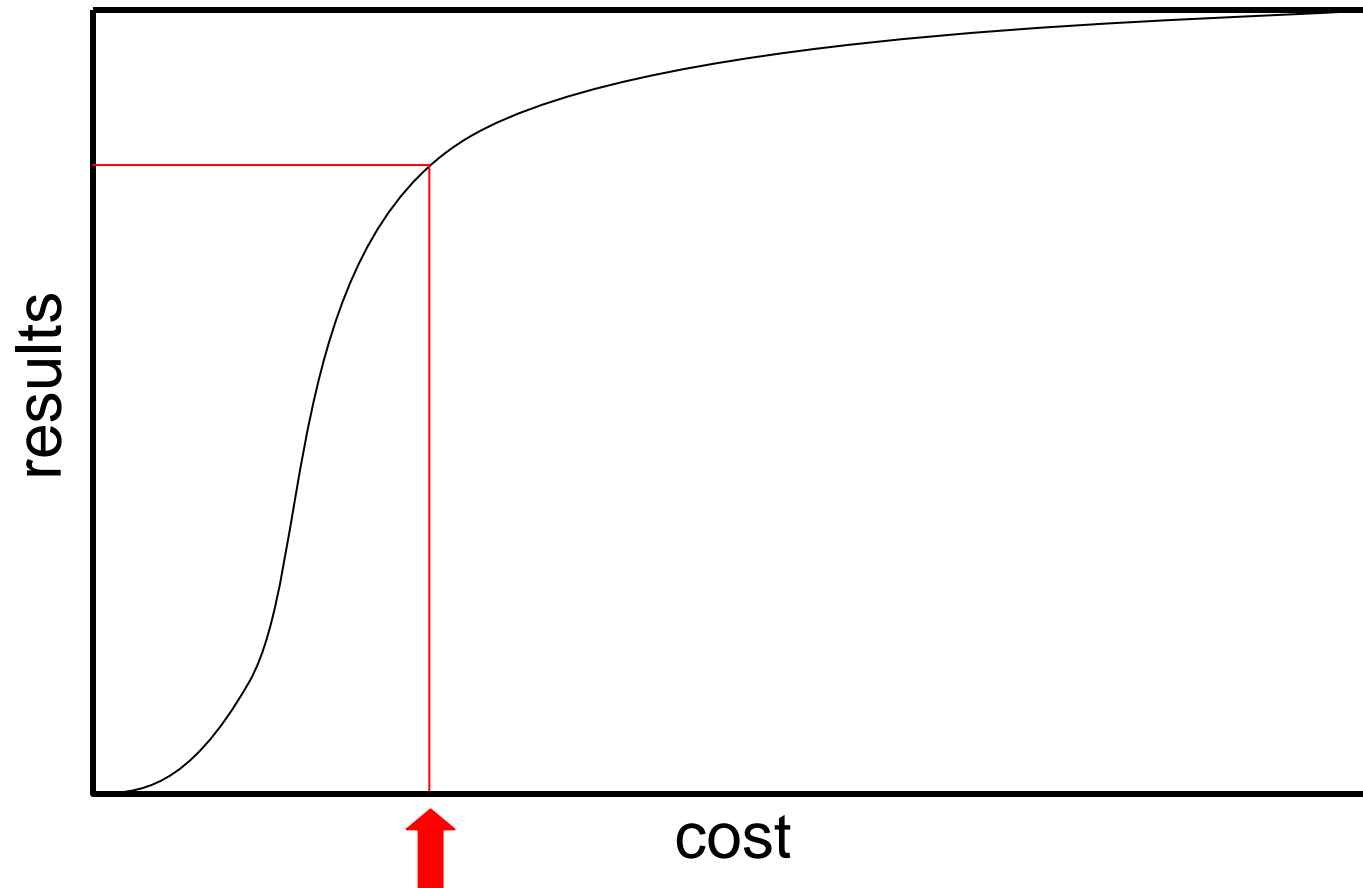
ECOLOGICAL RESEARCH FROM A TRULY LONG-TERM PERSPECTIVE:  
WHAT NEEDS TO BE DONE TODAY TO SUPPORT ECOLOGICAL ANALYSIS IN 2100

## The 80:20 Rule



ECOLOGICAL RESEARCH FROM A TRULY LONG-TERM PERSPECTIVE:  
WHAT NEEDS TO BE DONE TODAY TO SUPPORT ECOLOGICAL ANALYSIS IN 2100

## The 80:20 Rule



### ECOLOGICAL RESEARCH FROM A TRULY LONG-TERM PERSPECTIVE: WHAT NEEDS TO BE DONE TODAY TO SUPPORT ECOLOGICAL ANALYSIS IN 2100

#### 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.<sup>11</sup>



### ECOLOGICAL RESEARCH FROM A TRULY LONG-TERM PERSPECTIVE: WHAT NEEDS TO BE DONE TODAY TO SUPPORT ECOLOGICAL ANALYSIS IN 2100

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

*available*: the data must be 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.<sup>12</sup>




### ECOLOGICAL RESEARCH FROM A TRULY LONG-TERM PERSPECTIVE: WHAT NEEDS TO BE DONE TODAY TO SUPPORT ECOLOGICAL ANALYSIS IN 2100

#### 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<sup>14</sup> and the LTER Network has adopted a formal data-release policy:<sup>15</sup>

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:


*Metadata as a platform for choosing one or more be a diff*



### ECOLOGICAL RESEARCH FROM A TRULY LONG-TERM PERSPECTIVE: WHAT NEEDS TO BE DONE TODAY TO SUPPORT ECOLOGICAL ANALYSIS IN 2100



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.

ECOLOGICAL RESEARCH FROM A TRULY LONG-TERM PERSPECTIVE:  
WHAT NEEDS TO BE DONE TODAY TO SUPPORT ECOLOGICAL ANALYSIS IN 2100

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.

---

ECOLOGICAL RESEARCH FROM A TRULY LONG-TERM PERSPECTIVE:  
WHAT NEEDS TO BE DONE TODAY TO SUPPORT ECOLOGICAL ANALYSIS IN 2100

---

**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).





### ECOLOGICAL RESEARCH FROM A TRULY LONG-TERM PERSPECTIVE: WHAT NEEDS TO BE DONE TODAY TO SUPPORT ECOLOGICAL ANALYSIS IN 2100

#### 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.

### ECOLOGICAL RESEARCH FROM A TRULY LONG-TERM PERSPECTIVE: WHAT NEEDS TO BE DONE TODAY TO SUPPORT ECOLOGICAL ANALYSIS IN 2100

#### 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.

### ECOLOGICAL RESEARCH FROM A TRULY LONG-TERM PERSPECTIVE: WHAT NEEDS TO BE DONE TODAY TO SUPPORT ECOLOGICAL ANALYSIS IN 2100

In 2003 NSF released the Atkins Cyberinfrastructure report<sup>27</sup> 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.”





### ECOLOGICAL RESEARCH FROM A TRULY LONG-TERM PERSPECTIVE: WHAT NEEDS TO BE DONE TODAY TO SUPPORT ECOLOGICAL ANALYSIS IN 2100

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

---

ECOLOGICAL RESEARCH FROM A TRULY LONG-TERM PERSPECTIVE:  
WHAT NEEDS TO BE DONE TODAY TO SUPPORT ECOLOGICAL ANALYSIS IN 2100

P  
A  
S  
T  
A



---

ECOLOGICAL RESEARCH FROM A TRULY LONG-TERM PERSPECTIVE:  
WHAT NEEDS TO BE DONE TODAY TO SUPPORT ECOLOGICAL ANALYSIS IN 2100

P Provenance

A Aware

S Synthesis

T Tracking

A Architecture

ECOLOGICAL RESEARCH FROM A TRULY LONG-TERM PERSPECTIVE:  
WHAT NEEDS TO BE DONE TODAY TO SUPPORT ECOLOGICAL ANALYSIS IN 2100

P	Provenance	Full
A	Aware	Access
S	Synthesis	Research
T	Tracking	Information and
A	Architecture	Analysis
		Now

ECOLOGICAL RESEARCH FROM A TRULY LONG-TERM PERSPECTIVE:  
WHAT NEEDS TO BE DONE TODAY TO SUPPORT ECOLOGICAL ANALYSIS IN 2100

P Provenance

A Aware

S Synthesis

T Tracking

A Architecture

F Full

A Access

R Research

I Information and

A Analysis

N Now

---

ECOLOGICAL RESEARCH FROM A TRULY LONG-TERM PERSPECTIVE:  
WHAT NEEDS TO BE DONE TODAY TO SUPPORT ECOLOGICAL ANALYSIS IN 2100

F Full

S System

M Makeover

ECOLOGICAL RESEARCH FROM A TRULY LONG-TERM PERSPECTIVE:  
WHAT NEEDS TO BE DONE TODAY TO SUPPORT ECOLOGICAL ANALYSIS IN 2100



---

ECOLOGICAL RESEARCH FROM A TRULY LONG-TERM PERSPECTIVE:  
WHAT NEEDS TO BE DONE TODAY TO SUPPORT ECOLOGICAL ANALYSIS IN 2100

# END