Linguistic Data Consortium

Reproducible Research: Challenges for the Open Access movement

Mark Liberman

http://ling.upenn.edu/~myl

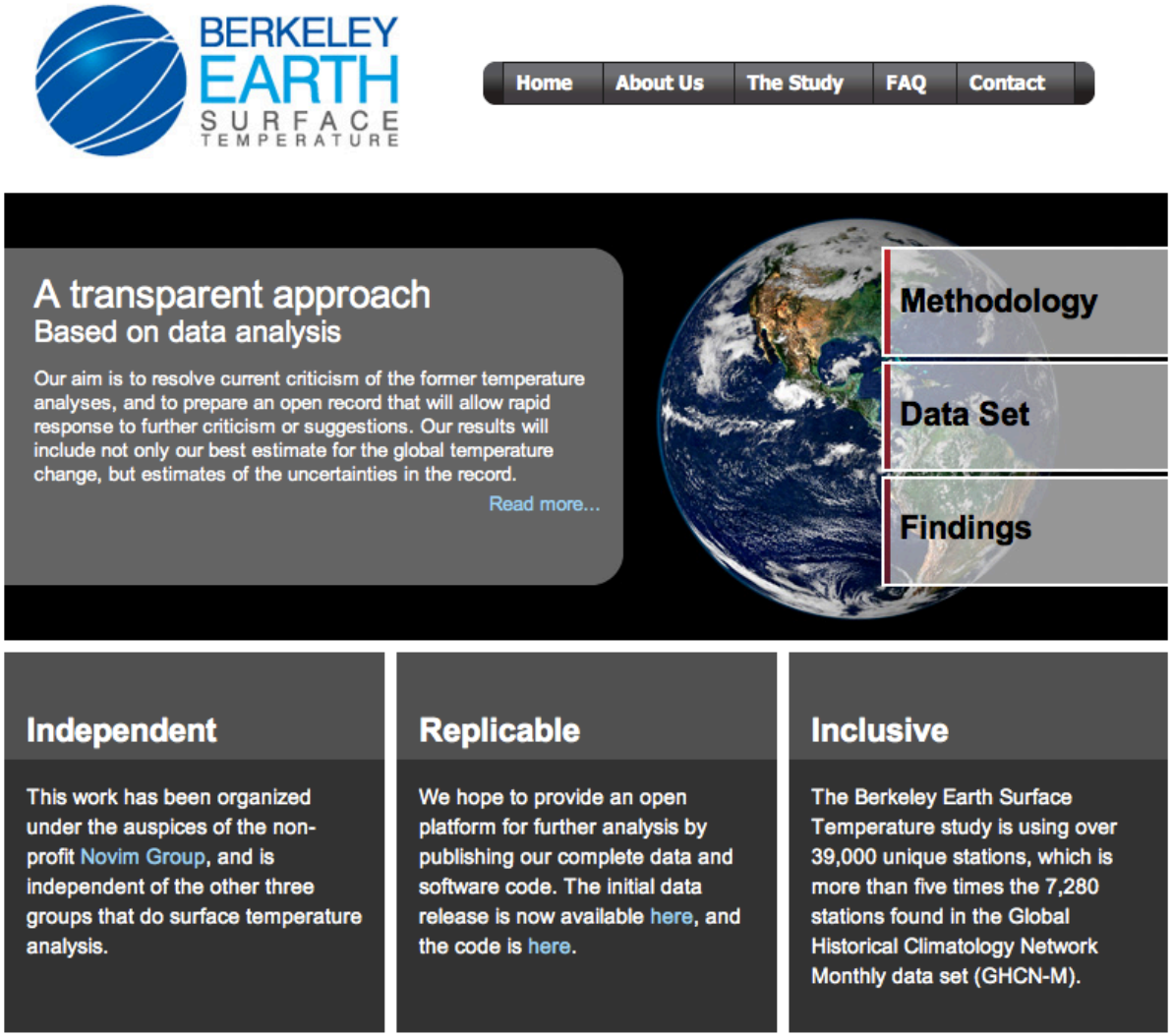
Penn
University of Pennsylvania

What is “Reproducible Research”?

Reproducible computational experiments,
where published source code
is run in a standard environment
on published data,
resulting in the numbers, tables, graphs etc.
that are discussed in a published paper.

Example: Berkeley Earth Surface Temperature Study

“We hope to provide an open platform for further analysis by publishing our complete data and software code.”


BERKELEY EARTH SURFACE TEMPERATURE
Home
About Us
The Study
FAQ
Contact
A transparent approach
Based on data analysis
Our aim is to resolve current criticism of the former temperature analyses, and to prepare an open record that will allow rapid response to further criticism or suggestions. Our results will include not only our best estimate for the global temperature change, but estimates of the uncertainties in the record.
Read more...
Methodology
Data Set
Findings
Independent
This work has been organized under the auspices of the non-profit Novim Group, and is independent of the other three groups that do surface temperature analysis.
Replicable
We hope to provide an open platform for further analysis by publishing our complete data and software code. The initial data release is now available here, and the code is here.
Inclusive
The Berkeley Earth Surface Temperature study is using over 39,000 unique stations, which is more than five times the 7,280 stations found in the Global Historical Climatology Network Monthly data set (GHCN-M).

How is RR done?

- Reproducible Research is a stool with three legs:

  (1) The input data;

  (2) The programs that implement an experiment;

  (3) An explanation of what the experiment does, why it matters, and what the results are.

- Traditional scientific publication provides leg (3) only
- With digital replication and transmission, publication of (1) and (2) has become technically feasible

Linguistic Data Consortium

Why bother?

- Reproducible Research is more efficient:
  - It lowers barriers to (re-)entry;
  - It makes replication and extension easier (including by the original author(s)!);
  - It encourages collaboration and pooling of data, and sharing/comparison of methods.
- Reproducible Research is more credible:
  - It documents progress over time;
  - It reduces mistakes (or makes them more likely to be found and fixed);
  - It prevents fiddling, cherry-picking, and outright fraud (or makes them more likely to be caught).

Geophysics: A success story

- Data sharing is widespread
  e.g. IRIS
  = "Incorporated Research Institutions for Seismology"
  "…archives and distributes data
  to support the seismological research community"

- Code publication is becoming more common
  e.g. via Madagascar

  "open-source software package for multidimensional data analysis
  and reproducible computational experiments […]
  for researchers … in geophysics and related fields"

RR as a management strategy

- DARPA's Human Language Technology program has used the "Common Task Method" since 1985:
  - Training and "development test" data is published via LDC;
  - Code for quantitative evaluation of results is published;
  - "Evaluation test" data is held back & used by NIST in periodic formal evaluations.
- Results:
  - Justification for steady funding before applications are possible;
  - Formation of new research communities; and therefore
  - Real progress in speech recognition, machine translation, information extraction ("text analytics"), speaker recognition, … and many other areas.
- Now adopted by funding agencies in other countries.

And don't forget those mistakes …

High-throughput biological assays let us ask very detailed questions about how diseases operate, and promise to let us personalize therapy. Data processing, however, is often not described well enough to allow for reproduction, leading to exercises in "forensic bioinformatics" where raw data and reported results are used to infer what the methods must have been. Unfortunately, poor documentation can shift from an inconvenience to an active danger when it obscures not just methods but errors. […]

One theme that emerges is that the most common errors are simple (e.g., row or column offsets); conversely, it is our experience that the most simple errors are common.

[Keith Baggerly, "The Importance of Reproducibility in High-Throughput Biology: Case Studies", AAAS 2010.]

Affinities… and not

- In “reproducible research”,
  data and code
  (which are normally proprietary or secret)
  are shared and extended
  by the members of an open research community.
- So “Reproducible Research is Open Research”.
- But RR is **not** necessarily “Open Access Research” –
  - The crucial thing is publication of data and code,
    not open access to data and code;
  - And there are barriers to open access in this area:
    - Familiar ones like money and IPR;
    - And new ones like privacy, confidentiality, HIPAA, etc.

RR & OA: Familiar Problems

- Money:
  - It's more expensive to create, maintain, & distribute data sets than to do the same thing for journal articles.
  - Reviewing code is difficult – and can be expensive.
  - Who pays?
    - As usual, some combination of sponsors, authors, users, etc.
    - But the problems are different from journal publication (mostly but not entirely because they're bigger problems)
- Intellectual property:
  - Some data sets have commercial IPR issues;
  - Tangled issues of legacy (often "orphaned") data (& code)
  - Solutions usually involve "user agreements" which may be "free as in beer" but not "free as in speech"

Linguistic Data Consortium

A new problem

- Privacy, confidentiality, “human subjects” protections:
  - These apply to biomedical data from human subjects, and to most data in the social and behavioral sciences
  - Such restrictions are imposed by
    - Government policies,
    - Data collection protocols (past and future),
    - Personal or professional ethics.
- Can we “publish” such data?
  - Usually, but the solutions require:
    - anonymization, which can be expensive;
    - user agreements, which limit openness of access.
- User restrictions range from minor to massive.
- This is an old problem in science but a new one for OA!

Linguistic Data Consortium

What we need

- A frank discussion of costs and business models:
  - The basic problems are the same as in journal publishing.
  - However, the scale is larger,
    and the details vary widely from field to field.
  - "Just put it up on the net" is an admirable goal,
    but usually not a valid solution.
- A re-definition of "Open Access"
  consistent with data privacy and confidentiality:
  - Data publication can preserve P & C
    via anonymization and agreements about storage and access;
  - These issues overlap with IPR solutions;
  - Is there an "Open Access" version? If so, what?

Where we're going

The benefits of "Reproducible Research"
to funders, researchers, teachers, students, and the public
are so great that the RR movement(s) will certainly spread.
(…even among academics,
the most culturally conservative tribe on the planet…)

The "Open Access" movement is spreading even faster.

Keeping these movements consistent
and moving forward
will take thought, experimentation, and action.