Ethics in Using Data; Citation to Data

IST400/600

Jian Qin

Ethical issues related to data management

- When to publish (too early or too late?)
- Whether to share data
- Whether journal should publish null results
- What constitutes plagiarism
- How to determine and adjudicate cases of data fraud
- Open access' threats to peer review

Code of Good Scientific Practice

- Supervision of the research worker in training
- Development of research protocols
- Documentation, storage, custody and sharing of the data, records and biological or chemical material resulting from the research work (see handout)
- Research projects sponsored by the healthcare industry or by any other profit-making institution
- Practice of publication
- Authorship of scientific work
- Practice of peer review

http://ethics.iit.edu/codes/coe/municipal.inst.med.research.html

Ethical principles fostered in data management

- Micro-ethics: Principles or values that scientists are taught
  - Validity
  - Transparency of methodology
  - Appropriate use of statistics
  - Data sharing

Ethical principles fostered in data management

- Macro-ethics:
  - Values relating to public policy and the larger society in the nation or the world, and
  - Responsibility of scientists and data managers to decide
  - Examples:
    - What data should be shared, at what cost, and to whom?
    - What mode of research financing and data sharing will maximize world health?
    - Save the oceans?
    - Foster world peace and prosperity?
    - Help control pollution?
    - Educate scientists and laymen about science?

A data sharing case: what should they do?

- Lucy Smith, has pioneered the use of a certain class of drugs to treat AIDS. At a conference, Jones visits a poster presented by Peter Smith, a student in Maxwell Montgomery's laboratory. Montgomery has just set up his lab with a small, 1-year New Investigator Award from a private foundation.
- Peter Smith is his first and only student right now, but Montgomery was hoping to attract new postdocs by announcing his and Smith's new discovery at the poster session.
- Their discovery is that one of the drugs in the class championed by Jones inactivates a particular protein called Abc1. This conclusion is based on a huge data set (including thousands of candidate proteins and chemical compounds) originally generated by Montgomery and later added to by Smith. Both also labored very hard to find the precise conditions for administering the drug and have produced the first known antibody to Abc1.
- Jones concludes that the Abc1-inactivator discovered by Smith and Montgomery could also be used to treat cancer, not just AIDS! Once back at her university, Jones e-mails Montgomery asking for the data set and the antibody.
- On seeing Jones's e-mail, Montgomery fears that Jones will scoop him with his own data and antibody. He also worries for Smith's thesis and career. If someone else publishes this work, Smith will likely be able to get his Ph.D., but he won't have a publication--not a good situation for a young scientist. Montgomery chooses to ignore Jones.
- But Montgomery seems to be ignoring Jones's e-mail and voice mail messages. Aggravated, Jones is considering putting pressure on Montgomery by contacting the head of his funding organization.

http://sciencecareers.sciencemag.org/career_development/previous_issues/articles/1680/sharing_in_the_sciences

The Political and Selective Use of Data: Cherry-Picking Climate Information in the White House

- The claim – the United States was doing better than the European Union in reducing greenhouse gas emissions
- The fact:

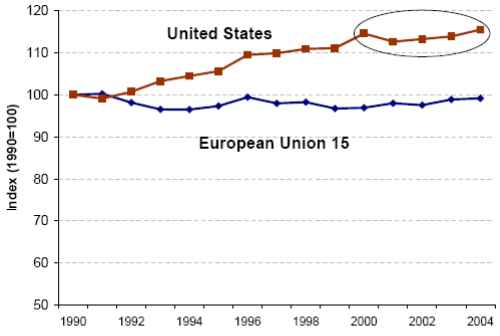

Figure 1. Index of Greenhouse Gas Emissions for the United States and the European Union, from 1990 to 2004. Index =100 for 1990. The artificial reference period selected by the White House is circled.[14]

The Political and Selective Use of Data: Cherry-Picking Climate Information in the White House

- Methods of selective use of data for unethical purposes:
  - Cherry picking of indicators
  - Cherry picking of time period
- What implications do these behaviors have on scientific data management?

Guidelines for

CITATIONS TO DATA

Based on Citing SEDAC data, applications and Web resources,
http://sedac.ciesin.columbia.edu/citations/CitationGuidelines.html

Why citations to data?

- Important for researchers, funding agencies, and data sources and partners to know that the data and information products we distribute are useful to the user community we support. One way of doing this is by tracking the use of data and information in publications.
- Important to acknowledge the authors and developers of a dataset, whether published or unpublished.
- Important to the integrity of science in that it provides the relevant information needed for readers to obtain a copy of the same data for further information or analysis.

Basic identifying information to be included in a citation

- Primary responsibility
- Title of the work
- Year of publication, issue, release
- Edition/version
- Type of resource, format
- Physical medium
- Statement of responsibility for dynamically generated data and maps
- Publisher and place of publication
- Distributor
- Availability and access
- Retrieval statement

Unpublished data (1)

Dataset is not part of a collection:

Contributing Authors. Pub Year. Dataset Title [format and/or medium]. Publisher location: Publisher Name. URL. Date accessed.

Example:

Center for International Earth Science Information Network (CIESIN), International Food Policy Research Institute (IFPRI) and World Resources Institute (WRI). 2000. Gridded Population of the World (GPW), Version 2 [online data]. Palisades, NY: CIESIN, Columbia University. Available at http://sedac.ciesin.columbia.edu/plue/gpw, retrieved July 1, 2003.

Unpublished data (2)

Dataset is part of a collection (Collection: Dataset):

Contributing Authors. Pub Year. Collection: Dataset Title [format and/or medium]. Publisher location: Publisher Name. URL. Date accessed.

Example:

Center for International Earth Science Information Network (CIESIN). 1996. Archive of Census Related Products (ACRP): 1990 Summary Tape file (STF1B) [online data]. Palisades, NY: CIESIN. Available at: http://sedac.ciesin.columbia.edu/plue/cenguide.html, retrieved July 1, 2003.

Unpublished data (3)

Dataset is part of a collection (Dataset. From: Collection):

Contributing Authors. Pub Year. Dataset Title [format and/or medium]. From: Collection. Publisher location: Publisher Name. URL. Date accessed.

Example:

Chinese Academy of Surveying and Mapping (CASM), University of Washington China in Time and Space (CITAS) and Center for International Earth Science Information Network (CIESIN) (1996). China Administrative Regions GIS Data: 1:1M, County Level, 1 July 1990. From: China Dimensions Data Collection. Palisades, NY: CIESIN. http://sedac.ciesin.columbia.edu/china/admin/bnd9071/bnd9071.html, retrieved July 1, 2003.

Published data

GENERAL FORMAT:

Contributing Authors. Pub Year. Title of the Work. Publisher location: Publisher Name. URL. Date accessed.

EXAMPLE:

Data published as a report. The report and data are available online:

World Economic Forum, Yale Center for Environmental Law and Policy, and CIESIN. 2002. Environmental Sustainability Index. New Haven, Ct.: Yale Center for Environmental Law and Policy. Available at: http://www.ciesin.columbia.edu/indicators/ESI/downloads.html; accessed July 1, 2003.

Summary

- Need to know ethics in data management and use at both micro- and macro-levels
- Data management to support the ethics requirements for
  - Validity
  - Reliability
  - Verification
  - Good science
- Citations are a way to acknowledge the rights holder and demonstrate uses of the data