The data flood: Implications for data stewardship and the culture of discovery

Margaret Leinen
Vice Chancellor for Marine Science, UC San Diego
Director, Scripps Institution of Oceanography

DataONE Webinar
September 8, 2015
observation (data) is the first step toward understanding
Five exabytes of content were created between the birth of the world and 2003. In 2013, five exabytes of content were created each day.

- ACI Information Group
Data growth is poised to exceed Moore’s Law growth: average growth = 64%, with many areas growing faster
CONSEQUENCES OF THE FLOOD
We’ve gotten used to not thinking about storage as a problem, but...
We’ve also gotten in the habit of thinking of data as the flood of automatically collected digital data, but...
What do we need to preserve and for how long?
What do we need to preserve and for how long?
Susan Solomon’s ‘The Coldest March’

• Combined
  – meteorological data
  – Logbooks and journals of Scott’s expedition
  – Current modeling and understanding
• Insights about the climatic forces influencing history

• We need to keep making this possible
THE IMPACT ON CULTURE OF SCIENCE
“The ultimate measure of success is the replicability of science”
- Coalition for Publishing Data in the Earth and Space Sciences (COPDESS)

But we deal with irreproducible events and observations
What does reproduceability mean in the geosciences with respect to data?
COPDESS Statement of Commitment

- Online community directory of appropriate community repositories
- Promulgate metadata information and domain standards
- Promote education of researchers in data management
- A working committee to update and curate this directory
- Reference data sets using the Joint Declaration of Data Citation Principles
- Include in research papers concise statements indicating where data reside
- Promote and implement links to data sets in publications and corresponding links to journals in data facilities
- Promote use of other relevant community permanent identifiers for samples (IGSN), researchers (ORCID), and funders and grants (FundRef).
- Develop workflows within the repositories that support the peer review process
Signatories

- American Astronomical Society
- American Geophysical Union
- American Meteorological Society
- Woods Hole Oceanographic Institution (BCO-DMO)
- Center for Open Science
- CLIVAR and Carbon Hydrographic Data Office
- COOPEUS
- Community Inventory of EarthCube Resources
- CUAHSI
- Continental Scientific Drilling Coordination Office
- Council of Data Facilities
- Dryad
- Elsevier
- European Geophysical Union
- Scripps Institution of Oceanography
- Springer
- ICSU World Data System
- Geological Society of London
- Geochemical Society
- Geological Society of America
- Incorporated Research Institutions for Seismology
- Integrated Earth Data Applications (IEDA)
- John Wiley and Sons
- LacCore: National Lacustrine Core Facility
- Magnetics Information Consortium (MagIC)
- Mineralogical Society of America
- Neotoma Paleoecology Database
- National Snow and Ice Data Center
- Nature Publishing Group
- OpenTopography
- Paleontological Society
- Proceedings of the National Academy of Sciences
- Rolling Deck to Repository (R2R) Program
- Science
- UNAVCO
Reproducibility in the field sciences

A May 2015 workshop sponsored by the Laura and John Arnold Foundation, AAAS/Science, American Geophysical Union, Ecological Society of America
Annotation and data
• Not interoperable
• Not open source
• Not standards based
• Not able to be used everywhere
• Not federated

<table>
<thead>
<tr>
<th>bookmarking</th>
<th>tagging</th>
<th>note taking</th>
<th>discussion</th>
</tr>
</thead>
<tbody>
<tr>
<td>xmarks</td>
<td>delicious</td>
<td>EVERNOTE</td>
<td>discussion</td>
</tr>
<tr>
<td>Google Chrome</td>
<td>Pinboard</td>
<td>Microsoft OneNote</td>
<td>Google +</td>
</tr>
<tr>
<td>Firefox</td>
<td>diigo</td>
<td>hello</td>
<td>Twitter</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Y Combinator</td>
</tr>
</tbody>
</table>
target address message target
http://domain/path
Annotations are like the web

Interoperable
Decentralized
Characterized by open source implementations
Other kinds of data
New trends merging ideas about data with those about samples

Challenges for universities and others who have sample collections
Whose responsibility is it to maintain archives?

Challenges for universities and others who have data archives and sample collections

Does the responsibility last forever?

Many universities are losing the ability to steward these collections/archives appropriately
Whose responsibility is it to maintain archives?

Government data archives

Already seeing cracks in the ability of government archives to morph to new technology, accommodate innovation in systems and structures
Data and scholarship
The emergence of doi’s making data sets referenceable

Peer reviewed publications emphasizing data sets

Challenges for scholarly advancement

Little consensus on when data sets constitute scholarship
Argo GDAC monthly snapshots DOIs

2015

➤ Snapshot of 2015 month 02
  http://dx.doi.org/10.12770/98ccefdc-39e4-4553-a0ed-6d16c2f1cbd1

➤ Snapshot of 2015 month 01
  http://dx.doi.org/10.12770/57ffbceb-bcb1-4295-a344-5e6c65151583

2014

➤ Snapshot of 2014 month 12
  http://dx.doi.org/10.12770/7693bf1c-7b95-484d-a55c-7453d54719ec

➤ Snapshot of 2014 month 11
  http://dx.doi.org/10.12770/99e07802-e8cb-48c8-a1f9-f2b6265c8870

➤ Snapshot of 2014 month 10
  http://dx.doi.org/10.12770/503f0ebf-2c58-41c7-81f5-a5993b4af9a3

➤ Snapshot of 2014 month 09
  http://dx.doi.org/10.12770/bc3de4fa-6668-4e0e-bae3-102c6d9c8ddd

➤ Snapshot of 2014 month 08
  http://dx.doi.org/10.12770/57b95b6a-ef27-47db-b14f-f8cb7c729793
NOAA Atlas NESDIS 72

WORLD OCEAN DATABASE 2013


Editor: Sydney Levitus
Technical Editor: Alexey Mishonov

National Oceanographic Data Center
Ocean Climate Laboratory

Silver Spring, MD
September 2013

U.S. DEPARTMENT OF COMMERCE
Penny S. Pritzker, Secretary

National Oceanic and Atmospheric Administration
Kathryn D. Sullivan,
Acting Under Secretary of Commerce for Oceans and Atmosphere

National Environmental Satellite, Data, and Information Service
Mary E. Kicza, Assistant Administrator
THINKING OF THE FUTURE
Brooks Hanson (AGU Director of Publications):

The publication of the future:

“I think the next big step, which the community is really trying to work on with some good initial steps but a long way to go is to really change what a publication is—to make data, code, etc. interoperable and seamlessly part of a publication so that a reader can move back and forth, rerun code; manipulate presentations of data sets.”
One of 1,000+ TED Talks

New ideas every weekday

TED.com
'Big Data' Skills Not Being Taught in K-12, Experts Say

By Benjamin Herold on December 17, 2014 6:08 AM