BCO-DMO

Biological and Chemical Oceanography Data Management Office

Introduction to Data Management for Ocean Science Research

Cyndy Chandler

Biological and Chemical Oceanography Data Management Office

EPOCA

European Project on OCean Acidification

12 November 2009
Ocean Acidification Short Course
Woods Hole, MA USA

NSF

Ocean Acidification

Studying ocean acidification's effects on marine ecoystems and biogeochemistry

Discussion Topics

Part 1 of 2: Introduction

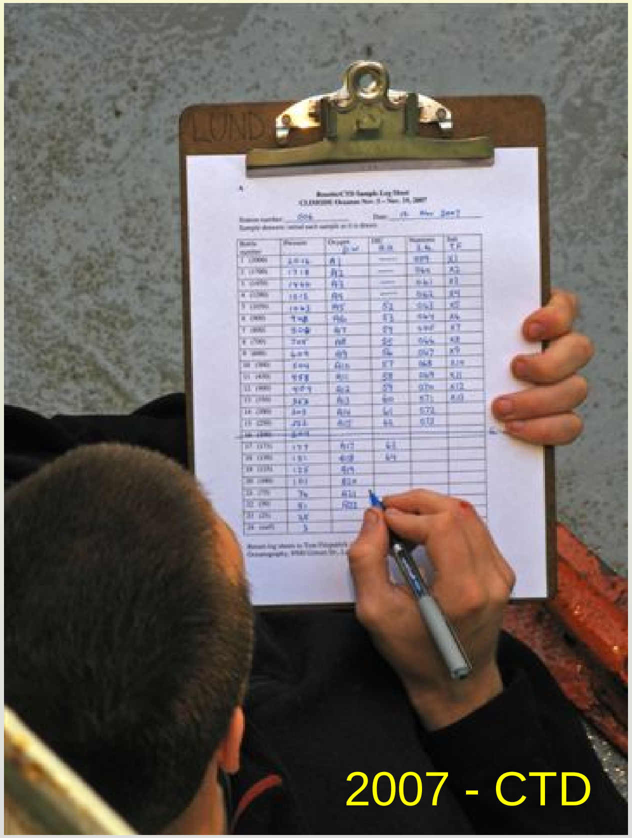
- Why data management matters
- New funding agency requirements
- New research paradigms
- New expectations for data access

Part 2: data management specifics

Why data management matters

- good data management practices have always been integral to the scientific method

1949 – recording BT



2007 - CTD

Why data management matters

It’s important to science

- careful and deliberate record keeping
- results reported and made publicly available
- enabling reproducibility of results

from the pre-course survey

- 57% of students reported having ‘minimal experience’ with “Metadata production and data archiving”

Some definitions … what do I mean by …

Data Management

- end-to-end data management
- proposal to preservation
- having a plan from the beginning
  to ensure that data and metadata
  are recorded accurately,
  are preserved securely (backups)
  and will be made accessible to others

and ‘dataset’ ?

- a logical grouping of related measurements
  (often from the same sampling device or sensor)

Metadata

- metadata ~ “about the data”
  information required to interpret the data

- Metadata records capture the information required to answer the who, what, where, why, how and when questions that are asked about a data set. It is important to know who collected, analyzed and contributed the data and where, when and how those data were acquired and subsequently analyzed and processed.

Changes and Challenges

- data sets used to be smaller
  and were often published on paper
  (in a journal article or a data report, and they fit in Table 1)
- data were published as a tangible thing
- as data acquisition becomes automated, rate of acquisition and volume increases
- but metadata acquisition (data documentation) is not being automated at the same rate

What else has changed?

- shift from ‘local’ to ‘global’
  - research themes
  - collaborative teams of researchers are trending toward being more distributed ~~ thematically and geographically
- technological advances are enabling these changes
- cultural changes lag behind technological changes
  - no direct relationship between career advancement and publication of data

Why data management matters

Cultural Changes – a work in progress:

- goal: scientific data should be freely accessible to all
- achievement of that goal relies on agreement that:

  anyone using the data
  must properly acknowledge the data originators
  (proper citation of all source data used)

Publication of Data

Cultural issues …

- little incentive for researchers to publish their data
- exacerbated by the perception that the data are the ‘property’ of the originating investigator, and might be ‘stolen’

Conventional wisdom is still that ‘publish or perish’ applies predominantly to journal publications, not data publication.

In the US, funding agency program managers are beginning to effect change in this area. NSF, NASA and NOAA all require publication of data generated by federally funded research.

New funding agency requirements

- Division of Ocean Sciences Data and Sample Policy. National Science Foundation. *NSF 04-004* http://*www.**nsf**.gov/pubs/2004/**nsf04004/nsf04004**.pdf*

  General Data Policy

  Principal Investigators are required to submit all environmental data collected to the designated National Data Centers as soon as possible, but no later than two (2) years after the data are collected. Inventories (metadata) of all marine environmental data collected should be submitted to the designated National Data Centers within sixty (60) days after the observational period/cruise.

New funding agency requirements

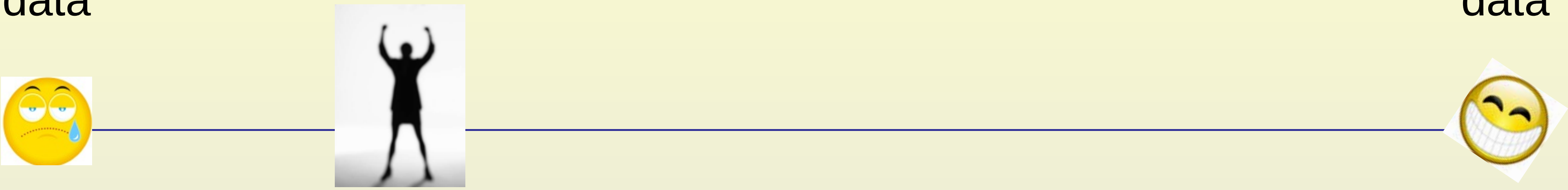
- **Proposal Requirements**
  The NSF Grant Proposal Guide requires that proposal Project Descriptions outline plans for preservation, documentation, and sharing of data, samples, physical collections, curriculum materials and other related research and education products. Plans for the handling of data and other products will be considered in the review process.

- **Reporting Requirements**
  Annual reports, required for all projects, should address progress on data and research product sharing. The Division of Ocean Sciences requires that final reports document compliance or explain why it did not occur.

Publication of Data

my
data

community
data



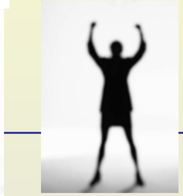
call me and I might share

freely available

Each approach has associated pros and cons, but as more data are published and are made freely available, it will become more of an accepted practice, and community expectations will change as well.

Paradigm Shift

Updating the ‘red phone paradigm’ . . .
developing new and better ways to locate and retrieve data.



familiar
it works

easy to learn
convenient
effective
yields better results

The grand challenge facing data managers today
is to design a data access system that can replace the telephone.

New research paradigms . . .

- science themes are trending toward
  - interdisciplinary
  - basin-wide

- studies involving coupling of complex models
  - atmospheric and hydrologic
  - end-to-end food web

. . . require access to data from many disciplines

New expectations for data access

- complex research themes (ocean biogeochemistry, ocean acidification research) require access to data collected by other researchers

- access to research designed to enable science-based decision support for legislative policies
  - social science
  - economics
  - history
  - broad range of disciplines

What does ‘access to data’ mean?

- ability to locate data of interest
- determine ‘fitness for purpose’
- accurately use the data

“Scientists are confronted with significant data management problems due to the large volume and high complexity of scientific data. In particular, the latter makes data integration a significant technical challenge.” (A.K. Sinha, Geoinformatics, 2006)

New expectations for data access

- New tools based on emerging technologies are being developed to address the challenge of

  integration of distributed heterogeneous data

- informatics
- semantic mediation
- registered ontologies

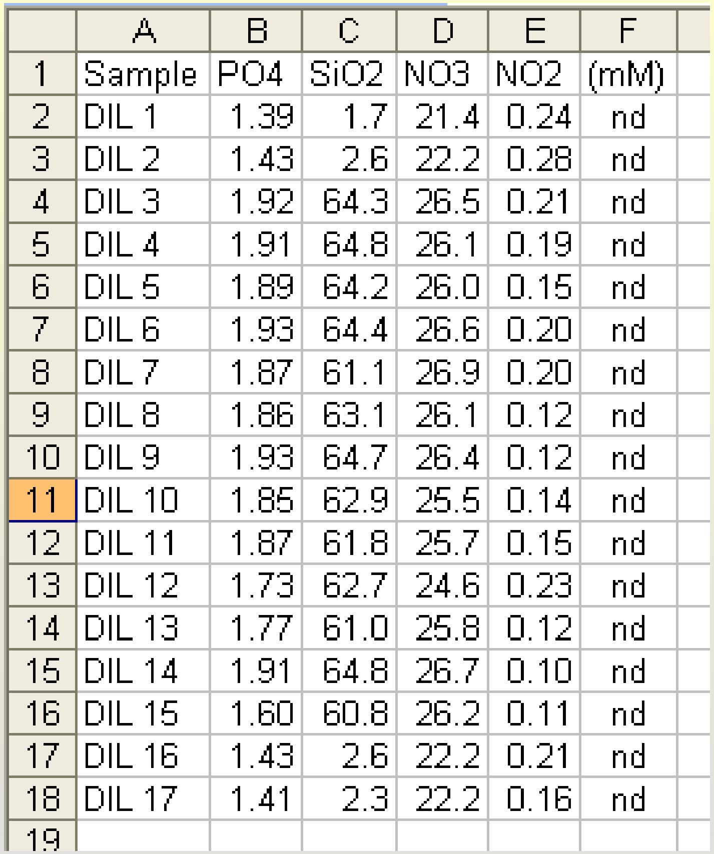
New expectations for data access

- all of the new technologies assume that data resources will be accompanied by machine-readable metadata

- while we wait for the new informatics tools, and semantic e-science resources to come online …

  … ocean science data accompanied by human readable metadata are of great value

these data . . .



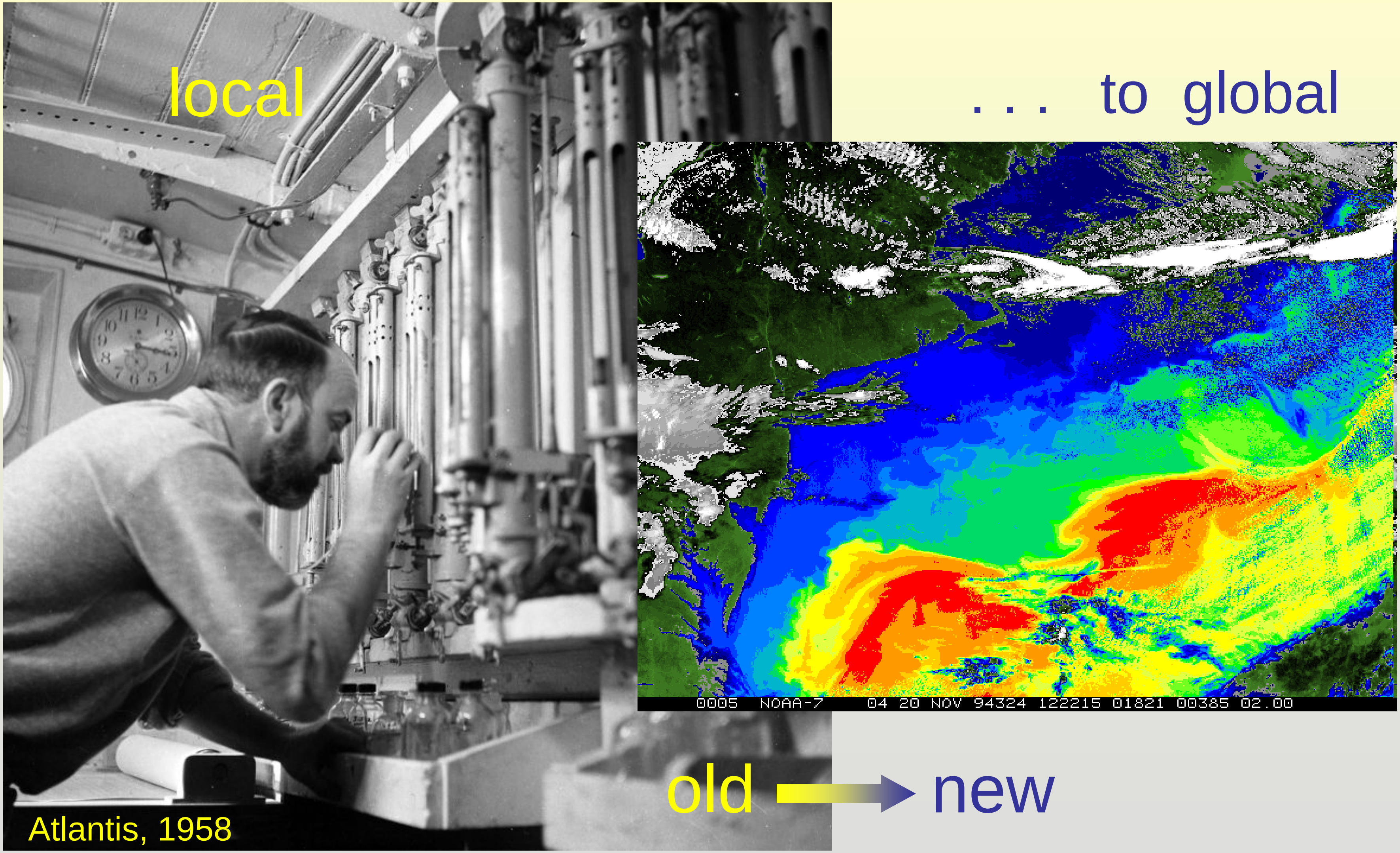
	A	B	C	D	E	F
1	Sample	PO4	SiO2	NO3	NO2	(mM)
2	DIL 1	1.39	1.7	21.4	0.24	nd
3	DIL 2	1.43	2.6	22.2	0.28	nd
4	DIL 3	1.92	64.3	26.5	0.21	nd
5	DIL 4	1.91	64.8	26.1	0.19	nd
6	DIL 5	1.89	64.2	26.0	0.15	nd
7	DIL 6	1.93	64.4	26.6	0.20	nd
8	DIL 7	1.87	61.1	26.9	0.20	nd
9	DIL 8	1.86	63.1	26.1	0.12	nd
10	DIL 9	1.93	64.7	26.4	0.12	nd
11	DIL 10	1.85	62.9	25.5	0.14	nd
12	DIL 11	1.87	61.8	25.7	0.15	nd
13	DIL 12	1.73	62.7	24.6	0.23	nd
14	DIL 13	1.77	61.0	25.8	0.12	nd
15	DIL 14	1.91	64.8	26.7	0.10	nd
16	DIL 15	1.60	60.8	26.2	0.11	nd
17	DIL 16	1.43	2.6	22.2	0.21	nd
18	DIL 17	1.41	2.3	22.2	0.16	nd
19						

. . . are incomplete and of little use to colleagues

The dataset lacks sufficient metadata to enable efficient and accurate reuse.

Presumably the data originator would decode Sample ‘DIL 10’ because they know it to be a proxy for where, when and how the data were collected.

Challenges and Opportunities



local

. . . to global

Atlantis, 1958

0005 NOAA-7   04 20 NOV 94324 122215 01821 00385 02.00

old → new

“ You can’t play with the data without the metadata.
Well, you can, but it’s much less fun. “
(Peter Wiebe, WHOI, 2009)

end of part 1