Opportunities and Challenges for Marine Data Management to Support Future Ocean Research

CYNTHIA CHANDLER

BIOLOGICAL AND CHEMICAL OCEANOGRAPHY DATA MANAGEMENT OFFICE

WOODS HOLE OCEANOGRAPHIC INSTITUTION

EMODnet Open Science Conference, Oostende, Belgium 20 October 2015

Outline

- Goals
- Challenges
- Opportunities
- Strategies
- Solutions
- How do we make sustainable progress?

Data Management

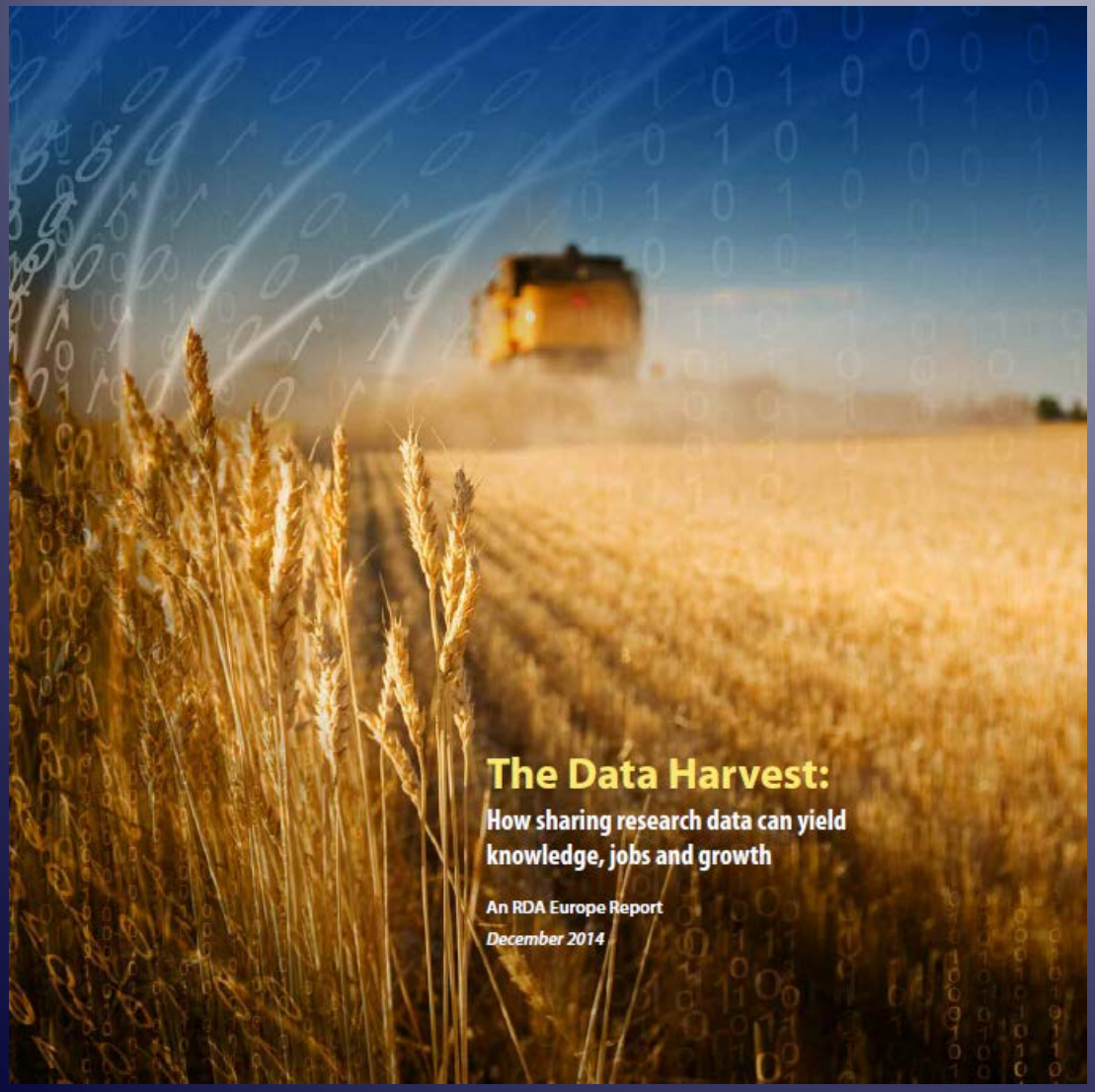
- Has always been essential to the scientific research process
- I have been managing field research data since I was 8 years old
- Much has changed in my lifetime with respect to data management
- Those changes represent challenges and opportunities for our community

CHANGING RESEARCH PARADIGM

- It is no longer enough to collect data by yourself, publish the paper and move on to the next research question.
- We see greater expectations from funding agencies, researchers, and the extended community for open data access and machine access.

D. McGuinness, Fall AGU 2012, Community Science - The Next Frontier

The Data Harvest



… a report published by Research Data Alliance – RDA Europe (December 2014)

https://europe.rd-alliance.org/

The Data Revolution

- “[The data revolution] isn’t just about the volume of scientific data; rather, it reflects a fundamental change in the way science is conducted, who does it, who pays for it and who benefits from it. And most importantly, the rising capacity to share all [these] data – electronically, efficiently, across borders and disciplines – magnifies the impact.”

The Data Harvest: How sharing research data can yield knowledge, jobs and growth (RDA Europe, December 2014)

Eight Global Sustainability Challenges

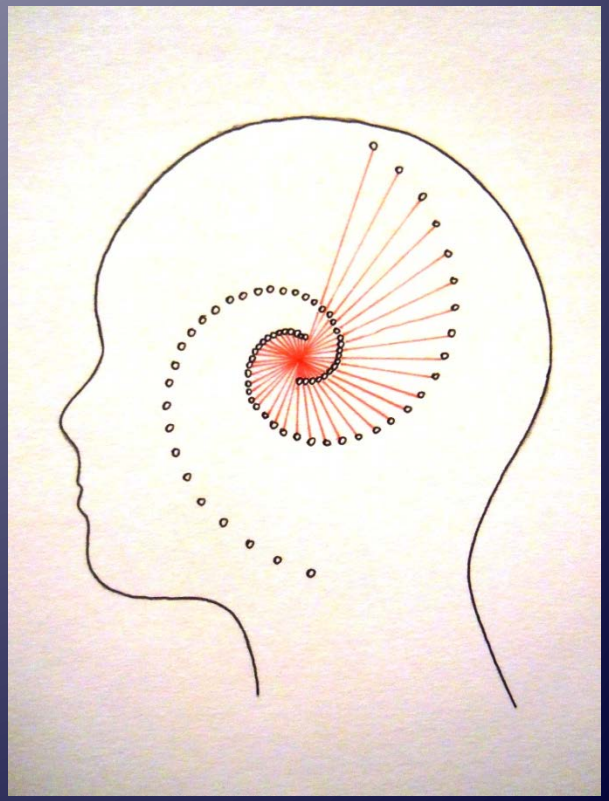
- Delivering water, energy, and food for all.
- Decoupling carbon emissions from economic growth.
- Safeguarding land, freshwater and marine natural assets.
- Building healthy, resilient and productive cities.
- Promoting sustainable rural futures.
- Improving human health by incorporating global change concerns.
- Encouraging sustainable consumption and production patterns.
- Improving governance and early warning systems to respond to complex future threats.

What do Researchers Need?

- Training to use new technologies
- Access to data and metadata (documentation)
- Access to near real-time data
- Cross-disciplinary research requires access to:
  - Discipline-specific metadata including quality control and provenance information
  - Published in machine-interpretable way
  - On-demand data visualization and integration

Challenge ~ Connectivity

- Goals & Strategies:
  - linking content curated at distributed repositories
  - improved interoperability (machine-to-machine)
- Technical solutions:
  - metadata content standards
  - controlled vocabularies
  - Linked Data, Brokering
- Not just technical
  - cultural conditions, behaviors
  - research data lifecycle
  - “proposal to preservation”

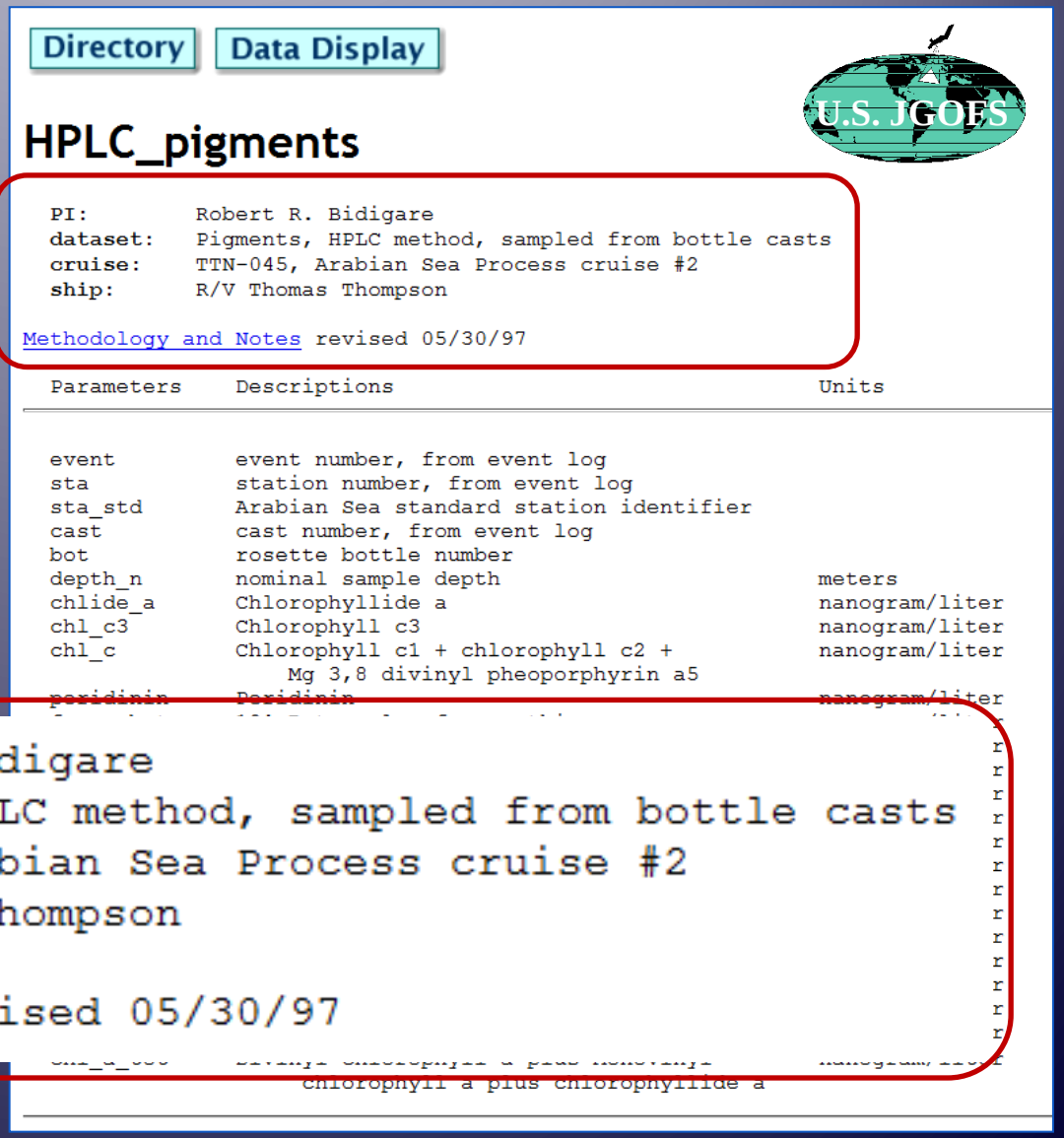


Strategy/Solutions: Connectivity

- Each ‘fact’ is explicitly declared and described by terms from (or linked to) community or global vocabularies
- Each term is identified by a globally unique Persistent IDentifier (PID)
- Each PID resolves to a semantic representation of that term, with relationships to other terms
- Published as open, standards-compliant records

Challenge #2 ~ Legacy Metadata

- Human readable text metadata records from environmental, legacy data systems



Directory | Data Display

U.S. JGOFS

HPLC_pigments

PI: Robert R. Bidigare
dataset: Pigments, HPLC method, sampled from bottle casts
cruise: TTN-045, Arabian Sea Process cruise #2
ship: R/V Thomas Thompson

Methodology and Notes revised 05/30/97

Parameters	Descriptions	Units
event	event number, from event log	
sta	station number, from event log	
sta_std	Arabian Sea standard station identifier	
cast	cast number, from event log	
bot	rosette bottle number	
depth_n	nominal sample depth	meters
chlide_a	Chlorophyllide a	nanogram/liter
chl_c3	Chlorophyll c3	nanogram/liter
chl_c	Chlorophyll c1 + chlorophyll c2 + Mg 3,8 divinyl pheoporphyrin a5	nanogram/liter
peridinin	Peridinin	nanogram/liter
	chlorophyll a plus chlorophyllide a	

PI: Robert R. Bidigare
dataset: Pigments, HPLC method, sampled from bottle casts
cruise: TTN-045, Arabian Sea Process cruise #2
ship: R/V Thomas Thompson

Methodology and Notes revised 05/30/97

Strategy: Metadata for Smart Data

- Use the **Semantic Web** to connect (or link) distributed data repositories
  - Formal ontology modeling
  - Create ontology design patterns
  - Use controlled vocabulary terms
  - Publish content as Linked Data

C. Bizer, T. Heath and T. Berners-Lee. 2009. "Linked Data - The Story So Far", International Journal on Semantic Web and Information Systems, Vol. 5(3), Pages 1-22. dx.doi.org/10.4018/jswis.2009081901

Solution: Metadata for Smart Data

- Publish as ISO 19139, W3C DCAT, schema.org Dataset extension
- Formal data publication with a DOI
- RDF with semantic markup including PROV

SHARED STANDARDS

Big Challenges & Big Opportunities

- Complex, large scale research questions
- Infrastructure (people, machines, systems) must be updated to support new research requirements
- Increased need for robust, discipline-specific, machine–actionable information (semantics)
- Standards-compliant metadata
- Documenting open-access data with PIDs (DOI)

Grand Challenge

- How to keep up with rapidly changing needs and expectations of
  - Research community
  - Funding agencies
  - Educators
  - Policy makers
  - Other stakeholders
- … given the usual limitations: funds, time, skilled personnel
- … and do it in a sustainable way?

Grand Challenge ~ Solution

- Strategic partnerships to develop, share & adopt common strategies and solutions
  - Domain-specific
  - Cross-domain
  - Regional
  - International

Grand Challenge ~ Solutions

COMMUNITIES OF PRACTICE

- US NSF EarthCube http://earthcube.org/
- ESIP http://esipfed.org
- Research Data Alliance https://rd-alliance.org/
- EMODnet http://www.emodnet.eu/
- ODIP http://www.odip.org/
- IODE http://iode.org/
- Future Earth http://www.futureearth.info/

BENEFIT FROM LONG-TERM COMMITMENT FROM AND ACTIVE ENGAGEMENT OF PROGRAM MANAGERS

"A scholar's positive contribution is measured by the sum of the original data that he contributes. Hypotheses come and go but data remain."

from: Advice to a Young Investigator (Santiago Ramón y Cajal, 1897)

Thank you

photo by Chris Linder (WHOI)

additional information

Woods Hole, Massachusetts, USA

- EarthCube
  http://earthcube.org/

  GOAL: to create integrated data management infrastructures across the US NSF geosciences by funding research efforts to complement, extend, enhance and connect existing infrastructure components.

ESIP

- Federation of
  Earth Science Information Partners
- http://esipfed.org/

  an open networked community that brings together science, data and information technology practitioners

Research Data Alliance

- https://rd-alliance.org/
- supported by the European Commission, the National Science Foundation and other U.S. agencies, and the Australian Government; constructing the social and technical bridges that enable open sharing of data across technologies and between disciplines and nations with the ultimate goal of addressing the grand challenges of society.

Ocean Data Interoperability Platform

http://www.odip.org/

Goal: contribute to the removal of barriers hindering the effective sharing of data across scientific domains and international boundaries. ODIP welcomes all the major organizations engaged in ocean data management in EU, US, and Australia and is supported by the IOC/IODE.

IODE

- International Oceanographic Data & Information Exchange
- http://www.iode.org/
- International community of domain-experts
- A network of National Data Centers and Associate Data Units
- Infrastructure already in place to support marine science research community

Belmont Forum

- Belmont Forum: http://www.bfe-inf.org/
- established in 2009, brings together environmental and geoscience funding agencies from 15 nations and seeks to build a coalition of national resources to advance global environmental change research.

ICSU
WORLD DATA SYSTEM

Future Earth

- http://www.futureearth.info/
- funding coordinated through the Belmont Forum, will be the platform through which many global change research programs will be coordinated, and the broad research themes, including the Earth Sciences, will require advanced information architectures to enable trans-disciplinary data-information-knowledge transfer