International Oceanographic Data and Information Exchange

Data Centre–Library Co-operation in Data Publication in Ocean Science

Roy Lowry, Gwen Moncoiffe and Adam Leadbetter (BODC)
Cathy Norton and Lisa Raymond (MBLWHOI Library)
Ed Urban (SCOR)
Peter Pissierssens (IODE Project Office)

International Council for Science
SCOR
Scientific Committee on Oceanic Research

Overview

- The players
- Meeting history
- Objectives and outcomes
- Data distribution paradigms and issues
- Pilot project activity
- Future plans

The Players

- Scientific Committee on Oceanic Research (SCOR)
  - International non-governmental organization
  - Formed by International Council of Scientific Unions (ICSU) in 1957
  - Scientists from 36 countries participate in working groups and steering committees
  - Promotes international cooperation
    - Planning and conducting oceanographic research
    - Solving methodological and conceptual problems that hinder research

The Players

- International Oceanographic Data and Information Exchange (IODE)
  - Data and metadata exchange programme of UNESCO Intergovernmental Oceanographic Commission (IOC)
  - Commenced in 1961
  - Modus operandi
    - Establish National Oceanographic Data Centres or Coordinators in IOC member states to acquire, enhance and exchange oceanographic data and metadata
    - Extend the NODC network through training and capacity building

The Players

- Marine Biological Laboratory Woods Hole Oceanographic Institution (MBLWHOI) Library
  - Woods Hole scientific community library with a strong interest in data publication in digital libraries
  - Digital Library Archive (DLA)
    - WHOI archives
    - Historical photographs and oceanographic instruments
    - Scientific data e.g. echo sounding records from WHOI research vessel expeditions
    - Technical report collections
    - Maps, nautical charts, geologic and bathymetric maps, and cruise tracks

Meetings to Date

- Ostend, June 2008 (http://www.iode.org/index.php?option=com_oe&task=viewEventRecord&eventID=273)
- Ostend, March 2009 (http://www.iode.org/index.php?option=com_oe&task=viewEventRecord&eventID=435)
- Jewett Foundation Woods Hole Data Repository Project Meeting, WHOI, April 2009 (http://tw.rpi.edu/portal/Jewett_Meeting_at_MBL)
- Paris, April 2010 (http://www.iode.org/index.php?option=com_oe&task=viewEventRecord&eventID=625)

Objectives and Outcomes

- Objectives
  - Engage the IODE Data Centre Community in data publication
  - Provide a network of hosts for cited data
  - Motivate scientists through reward for depositing data in data centres
  - Promotion of scientific clarity and re-use of data

Objectives and Outcomes

- Ostend 2008 Meeting
  - ‘Show and tell’
    - Digital repository perspective
    - Pangaea perspective
    - Data mining perspective
    - Digital library perspective
    - Editor perspective
    - Publisher perspective
    - Library perspective
    - Scientist perspective
  - A lot of information was shared but no clear way forward emerged

Objectives and Outcomes

- Ostend 2009 Meeting
  - Beginning to tease out use cases from a data centre perspective
    - Data publication service for accessions
    - Publication of enhanced data exported from data centre holdings
    - Accessibility of figure and table datasets ('numbers behind the graph')
  - Only the first two likely enhance the servable holdings of a data centre such as BODC
  - Clear role for digital libraries in the third

Objectives and Outcomes

- WHOI meeting
  - Pilot project agreed to test out data publication procedures using a Peter Wiebe publication and the MBLWHOI Library digital repository
  - Paper was a review paper with a huge number of potential data citations and therefore published before the citations and datasets could be assembled.
  - Further project based on a simpler paper

Objectives and Outcomes

- Paris 2010 Meeting
  - Understanding of the difference between ‘best available’ data serving and data publishing
  - Decision to directly contact IODE data centres to get them involved in data publication
    - Enthusiastic response: everybody wants to do it but nobody seems to know where to start
    - BODC has been trail blazing
    - Motivation attempt at IODE XXI meeting in Liege next March

Data Distribution Paradigms

- IODE Data Centre Paradigm
  - Data change significantly at the data centre
    - Value added to data through:
      - Metadata generation
      - Quality control (flagging outliers)
      - Raw data work-up (conversion of raw voltages to usable units followed by calibration against sample data)
      - Ingestion into a common schema (reformatting, relational database schema population)

Data Distribution Paradigms

- IODE Data Centre Paradigm
  - 'Best available' data served during data evolution
  - Change is continuous with no snapshots preserved or formal versioning during work-up
  - Data considered 'completed' may still change
    - Usage metadata continually improving
    - Additional quality control based on user feedback

Data Distribution Paradigms

- Digital Library Paradigm
  - Dataset is a ‘bucket of bytes’ which is
    - Fixed (checksum should be a metadata item)
      - Changes generate a new ‘version’ (snapshot with its own identifier and citation)
      - Previous versions must persist
    - Accessible on-line via a permanent identifier
    - Usable on a decadal timescale (standards e.g. OAIS)
    - Citable in the scientific literature
    - Discoverable

Data Distribution Paradigms

- Digital Library Paradigm
  - Technologies such as DSpace
    - Generates provenance information on ingestion
      - File size
      - Date
      - Depositor name
      - Checksum
    - Serves out exactly what is ingested
    - Supports a strategy where any data change causes
      - New dataset
      - New metadata
      - New DOI

Data Distribution Paradigms

- Digital Library Paradigm
  - Metadata founded on Dublin Core
    - Supports basic discovery
    - Insufficient for scientific discovery facets
      - Reinforce using standards such as IOS19115, DIF, FGDC, Darwin Core
    - Totally inadequate for scientific browse and usage
      - Reinforce using:
        - Text documentation in formats like PDF
        - Standards like SensorML and Observations and Measurements
  - Dublin Core provides an essential link to digital libraries and should not be ignored by data centres

Data Distribution Paradigms

- Mapping Data Centre to Digital Library
  - Science community pressures for BODC to engage in the Digital Library paradigm
  - Initial strategy
    - Identify data objects that could be tagged by DOIs
    - Use existing infrastructure to provide on-line access to objects
    - Dead end as no version management infrastructure

Data Distribution Paradigms

- Mapping Data Centre to Digital Library
  - BODC published 11 CD-ROM datasets from 1992-2001
    - Snapshot of value-added data exported from dynamic system
    - Perfect fit to the Digital Library paradigm
  - Could this process be updated and resurrected?
    - Snapshots of value-added data served as digital library objects
  - Issues
    - Time taken for adding value can exceed scientists' patience threshold
    - Where to publish the snapshot?

Data Distribution Paradigms

- Mapping Data Centre to Digital Library
  - Short turnaround data publication service also needed
    - Provide through extension to existing accession procedures
      - Specify standards for data submissions
        - Content, format, metadata, etc.
      - Check submissions against these standards
        - Pass could be part of a data publication editorial process
      - Tag with a URN (DOI is the obvious candidate)
      - Publish in a suitable repository
      - Post metadata with DOI binding
      - Generate Dublin Core metadata and citation

BODC Pilot Project Activity

- Ingested Dataset Publication
  - Marine and Freshwater Microbial Biodiversity dataset prepared for CD-ROM but never published
  - Project is to publish it which involves:
    - Dataset update and completion (done)
    - Submission to a suitable digital repository (under negotiation)
    - Obtaining a DOI (registered with British Library and DataCite)
    - Discovery metadata creation (ISO19139 done)

BODC Pilot Project Activity

- Establish Data Publication Service
  - Draft dataset guidelines summary
    - Data format
      - Technologically stable (e.g. ASCII, NetCDF)
      - International content standard (e.g. CF, SeaDataNet) conformant
      - Semantic layer uses established vocabularies (e.g. CF Standard Names, BODC PUV)
    - Data to be accompanied by metadata
      - Dublin Core
      - Enhanced discovery metadata (ISO19115, DIF, FGDC)
      - Usage metadata (HTML, PDF, SensorML, O&M)
    - Guidelines require review and development

Future Plans

- Completion of the pilot projects at BODC
- Raise awareness of our activities at AGU and IODE 50
- Engage other data centres in data publication through reporting the BODC activity and experiences
- MBLWHOI Library and BCO-DMO data centre data publication collaboration
- Meeting between University of Delaware College of Earth, Ocean and Environment and university library to set up a data publication pilot project

That's All Folks

- Thank you for your attention
- Questions?