IODE International Oceanographic Data and Information Exchange

MBLWHOI LIBRARY

Data Publication: A New Paradigm for IODE Data Management?

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

International Council for Science
SCOR
Scientific Committee on Oceanic Research

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)

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 Publication 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 Publication Paradigm

- Technologies such as D-Space
  - Serves out exactly what is ingested
  - Supports a strategy where any data change requires a new dataset, new metadata and a new DOI
- Metadata founded on Dublin Core
  - Supports basic discovery but insufficient for scientific discovery facets
    - Reinforce using standards such as IOS19115, DIF, FGDC, Darwin Core
  - Totally inadequate for scientific browse and usage
    - Reinforce using plaintext documentation or standards like SensorML and Observations and Measurements
- Dublin Core provides an essential link to digital libraries and should not be ignored by data centres

Paradigm Mapping

- IODE data centres produced CD-ROMs in the 1990s
  - Snapshot of value-added data exported from dynamic system
  - Perfect fit to the Digital Library paradigm
- Could this process be updated and resurrected with 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?

Paradigm Mapping

- Short turnaround data publication service also needed
  - Provide through extension to data centre 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 DOI
    - Publish in a suitable repository
    - Post metadata with DOI binding
    - Generate Dublin Core metadata and citation

Paradigm Mapping

- SCOR, IODE, MBLWHOI Library group set up in 2008 with the following objectives
  - Engage the IODE Community in data publication
  - Provide a network of hosts for cited data
  - Motivate scientists through reward for depositing data in data centres
  - Promote scientific clarity and re-use of data
- 4 meetings since June 2008
- CODATA Task Team created in October 2010 and currently spinning up activity
- Cross-communication between groups is in place

Pilot Project Activity

- IODE currently setting up the ‘Published Ocean Data’ D-Space repository
  - Parallel resource to OceanDocs
  - Designed to support pilot data publication projects

- BODC currently developing two pilot projects that will use this in conjunctions with DataCite DOIs via the British Library

Pilot Project Activity

- BODC pilot project work
  - 21st Century CD-ROM
    - Publish Marine and Freshwater Microbial Biodiversity dataset prepared for CD-ROM but never published
  - Data Publication Service
    - Data conforming to BODC-specified quality standards will be published prior to data centre ingestion
    - Designed to provide data citations in time for inclusion in published manuscripts

Pilot Project Activity

- MBLWHOI Library working with the BCO-DMO data centre at WHOI to publish master dataset accessions and cite them using DOIs

- University of Delaware College of Earth, Ocean and Environment and university library working together on a data publication pilot project

A Scalable Future

- Production data publication potentially requires access to multiple versions of large numbers of datasets
- The ‘Published Ocean Data’ repository cannot be expected to support this
- Data centres could establish a network of D-Space repositories for version snapshot storage
- Could work but it would force duplicated storage of multiple copies of datasets

A Scalable Future

- Can we be smarter?
  - Formal quantised versioning in data centre practices
  - Pragmatic review of dataset definitions
  - Dynamic recreation of past versions
    - Database rollback from change logs
    - Data file update through change scripts
    - Technologies from the computer science digital curation community
      - Workflows
      - Provenance metadata

Data Publication Conclusions

- ‘Citable datasets’ are digital extensions to published papers and so must be static
- Data publication benefits data providers, data centres and data users whilst providing transparency that upholds scientific integrity
- Demand for data publication from the scientific community is high and growing fast giving it a clear role in the future of data management
- Data centres that do not engage will be viewed as dinosaurs and could share their fate

That’s All Folks

- Thank you for your attention

- Questions?