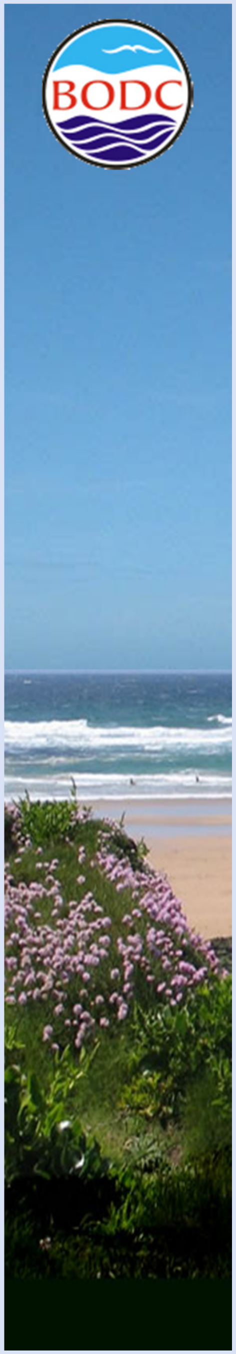
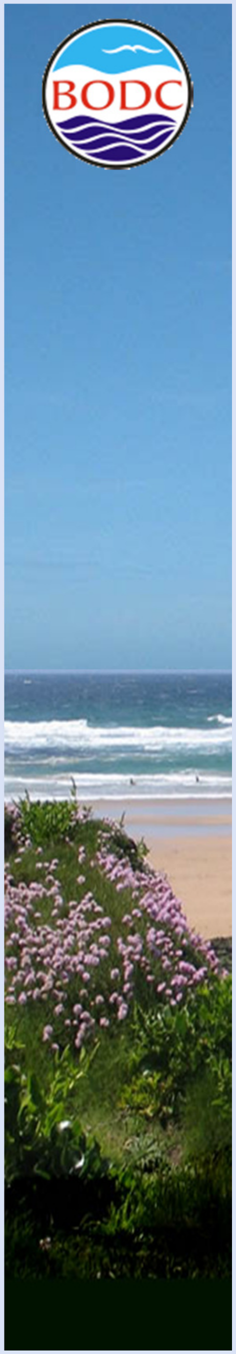


Mapping the data publication paradigm onto the operations of the British Oceanographic Data Centre

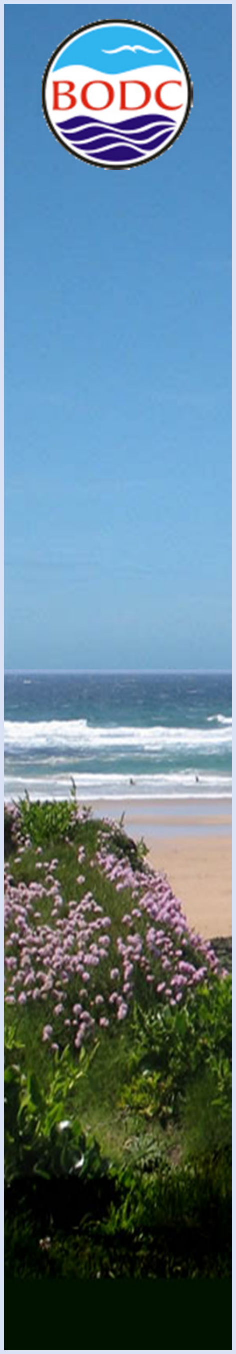
Roy Lowry

British Oceanographic Data Centre



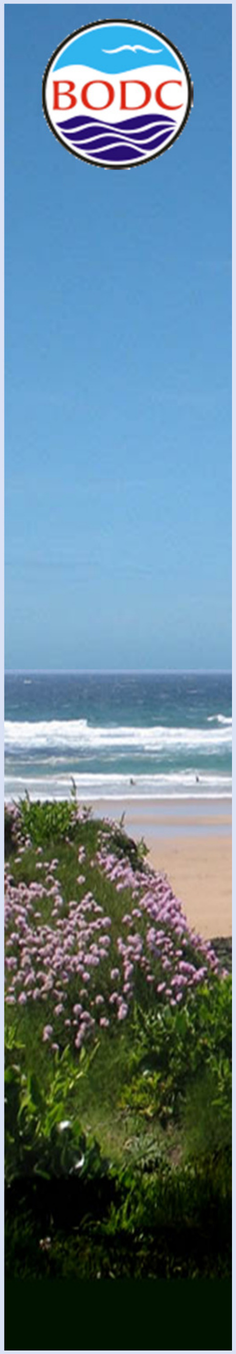
Summary

- Historical Perspective
- The IODE Data Centre Paradigm
- Data Publication Paradigm
- Paradigm Mapping
- Bringing Data Publication to BODC



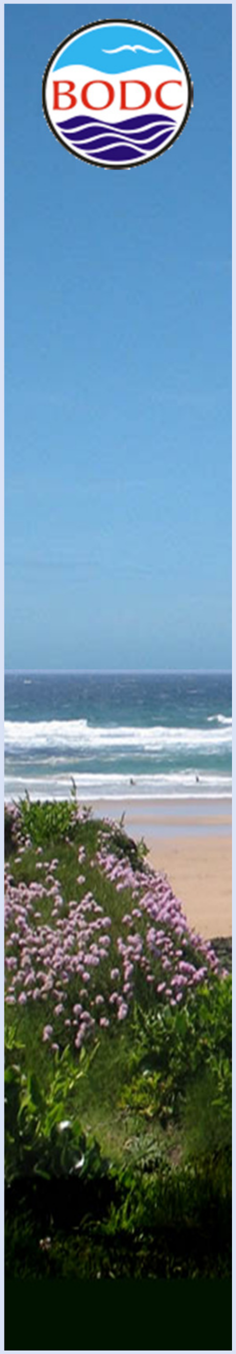
Historical Perspective

- **Oceanographic data are:**
  - **Extremely sparse**
    - Satellites cover surface not depth
    - Moorings cover a single point in space over time
    - Vessels and modern platforms (floats, gliders, AUVs) occupy a single trajectory through space and time
  - **Extremely expensive**
    - Research vessels cost millions to build and thousands of pounds per day to run
    - A small fortune in hardware is thrown into the sea and hopefully recovered at a later date



Historical Perspective

- This has resulted in a long-established data sharing culture in the oceanographic community
- IOC established the International Oceanographic Data and Information Exchange programme in 1961
- Fifty years on and it’s still going strong

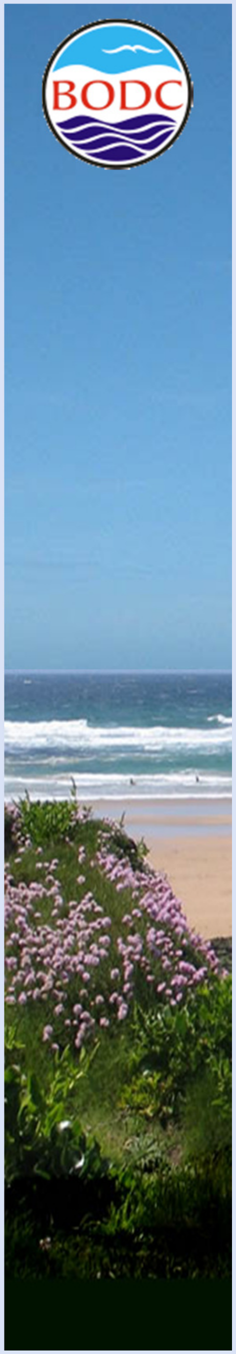


Historical Perspective

- IODE infrastructure in the form of National Oceanographic Data Centres was established in the 1970s and early 1980s

- Operational model developed through the activities of IODE's technical development group (GETADE)

- The resulting operational paradigm has been running in at least a dozen centres around the world for over 30 years



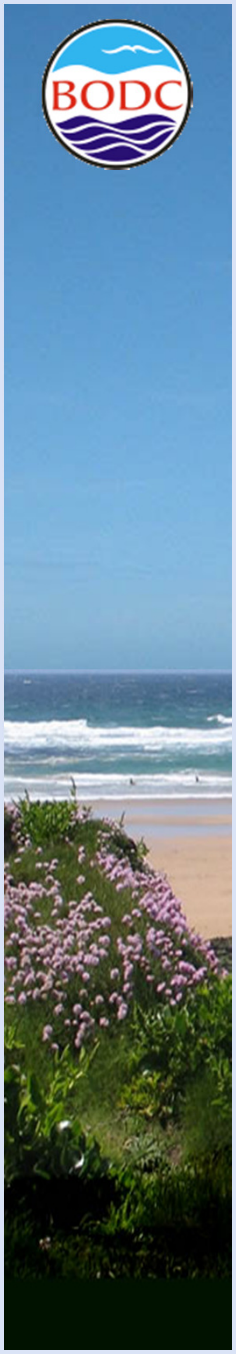
Historical Perspective

➢ A long history has its advantages

- Significant data holdings
- Experience leading to deep understanding of the data

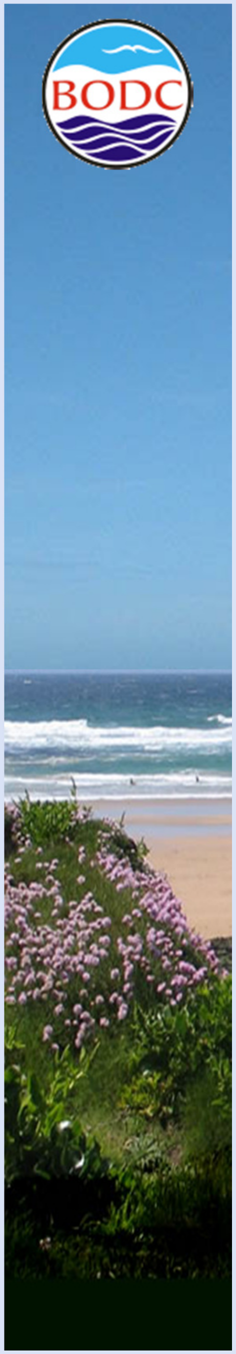
➢ But it also has its disadvantages

- Large amounts of legacy incorporating everything we did when we didn’t know better
- Massive inertia to be overcome by any change or adoption of new technology
- BODC’s migration from plaintext name and address metadata to ISO standards is a case in point



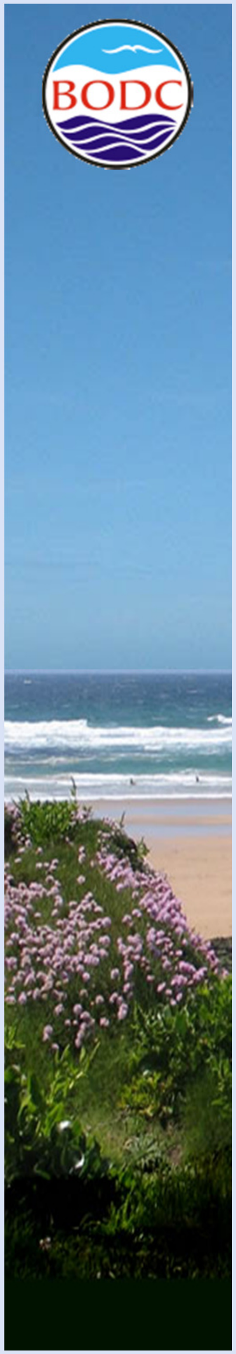
The IODE Data Centre Paradigm

- ➢ Data change significantly at the data centre
- ➢ Value is added to data through:
  - Metadata generation
    - ∗ Preparation of usage metadata by collation from logs, reports, papers etc.
    - ∗ Preparation of standard discovery metadata
    - ∗ Standardisation of the semantic layer
  - Quality control
    - ∗ Flagging outliers
    - ∗ Adding issue descriptions to usage metadata



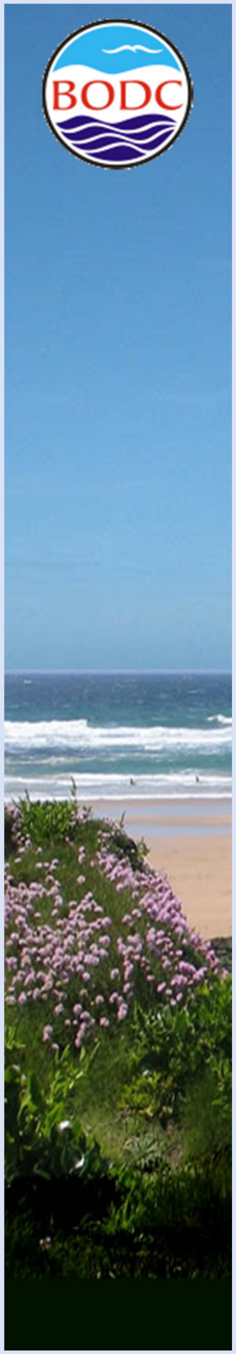
The IODE Data Centre Paradigm

- **Data change significantly at the data centre**
  - **Raw data get worked up**
    - ∗ Voltages and ADC counts converted to engineering units
    - ∗ Calibration against sample data
  - **Harmonisation through ingestion**
    - ∗ Reformatting into a uniform file format
    - ∗ Loading into a common RDBMS schema
  - **The result is a soup of data atoms bearing little resemblance to what was delivered**



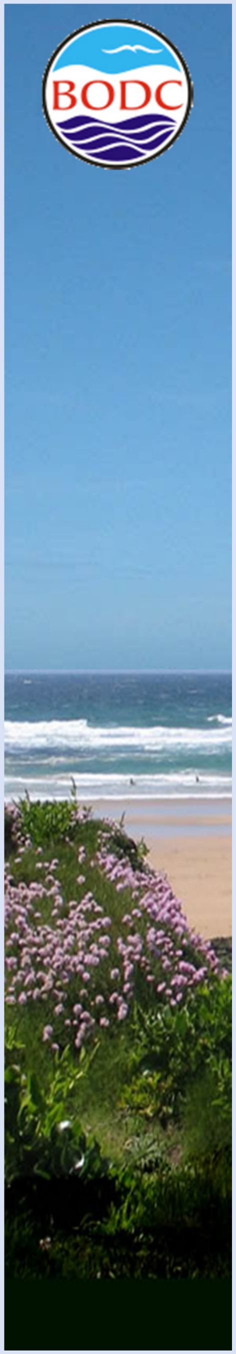
The IODE Data Centre Paradigm

- Service designed around supporting a ‘data synthesis input’ use case
- Data synthesis considered as a buffer between data centre output and scientific interpretation/publication
- ‘Best available’ data at the time of the request is served
- Change is continuous with no snapshots preserved or versioned checkpoints in the workflow



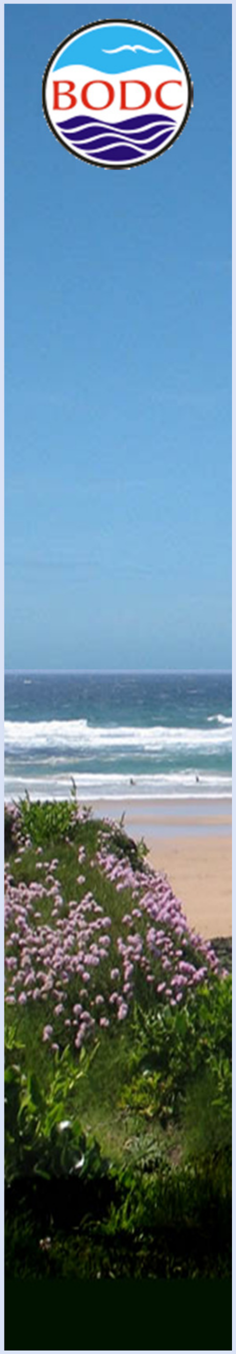
Data Publication Paradigm

- **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**



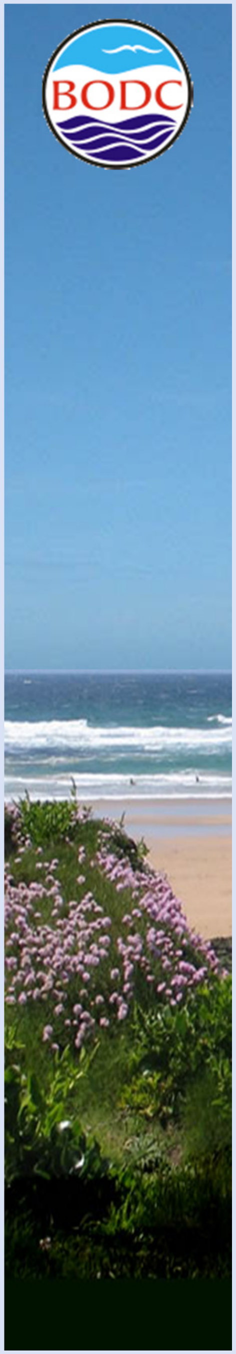
Data Publication Paradigm

- **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**
    - May be reinforced using plaintext documentation or standards like SensorML and Observations and Measurements



Paradigm Mapping Issues

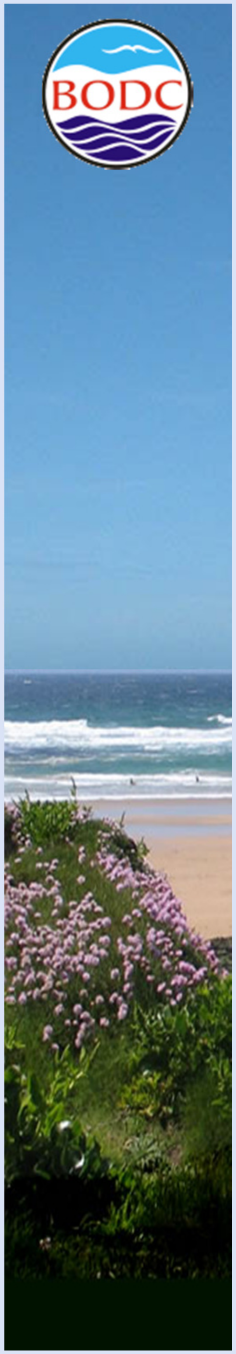
- **What is a dataset?**
  - Dynamic entity in the data centre paradigm that needs pinning down if it is to map to its static equivalent in publication
- **How can replicated serving be guaranteed?**
  - Migration from fluid change to a workflow based on quantised steps
  - Storage management and access to past versions
- **How can incompatibilities in workflow timing requirements be resolved?**
  - Data centre procedures add value to data but take a considerable length of time
  - Publication process wouldn’t welcome this as a blocker in their workflow
  - Requirement to provide permanent identifiers for vapourware datasets



Paradigm Mapping Solutions

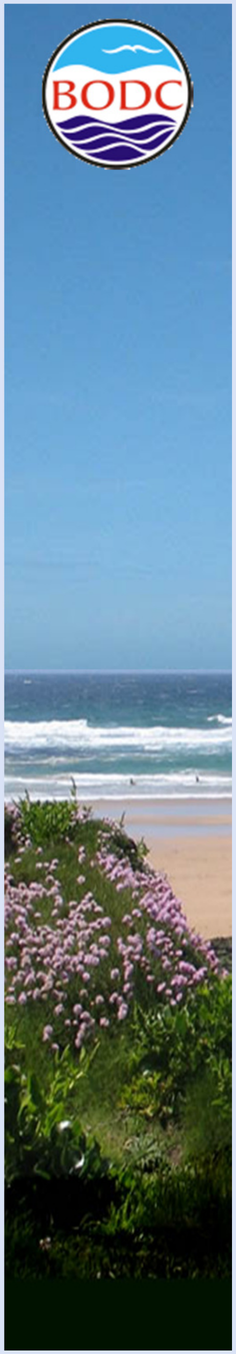
➢ What is a dataset?

- **Introduction of the ‘discovery dataset’ concept**
  - ∗ Systematic groupings of data atoms
  - ∗ Existed for decades but never closely coupled to data (EDMED legacy)
  - ∗ Programme underway in BODC put this right by physically mapping discovery datasets to their component data atoms
- **Introduction of ‘request publication’ concept**
  - ∗ Give user the option of publication when they create and download a dynamic dataset
  - ∗ Providing, of course, they supply the metadata required for DOI minting and landing page population!



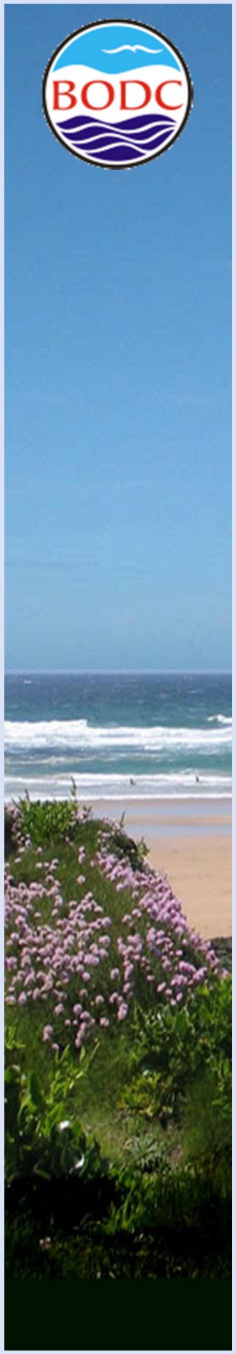
Paradigm Mapping Solutions

- How can replicated serving be guaranteed?
  - Introduction of ‘publication’ concept into ingestion workflows
  - Physical instantiation of usage metadata (currently a dynamic report served through a Web Service)
  - Introduction of past version storage management and access infrastructure



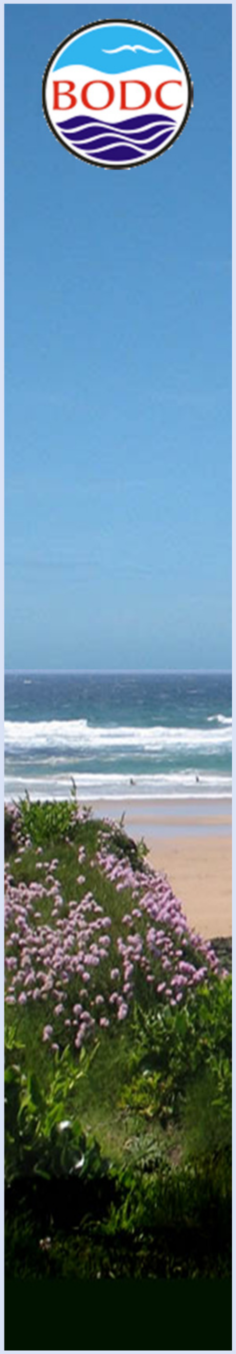
Paradigm Mapping Solutions

- ➢ **Addressing timing mismatches**
  - **Publication without ingestion**
    - ∗ Provide an accession publication service
      - – Accession dataset comes through the door
      - – Verified as up to scratch
      - – Placed in web-accessible storage
      - – DOI minted and data files linked to landing page
    - ∗ Caveats
      - – Data supplied must be data that BODC wish to ingest
      - – Precise definition of ‘up to scratch’ required
      - – Business scope expansion into ‘data behind the graph’ cannot be supported



Paradigm Mapping Solutions

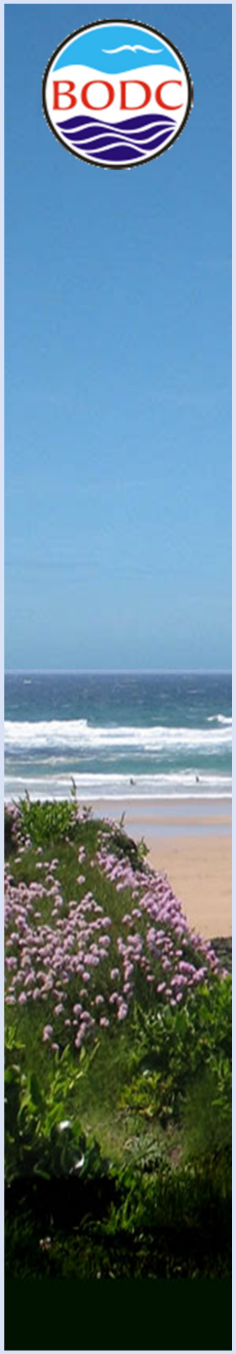
- **Addressing timing mismatches**
  - ∗ Expectation management
    - –Publishing promises will NEVER be considered by BODC
    - –No DOI will be minted without files verified as acceptable quality in BODC’s possession



Bringing Data Publication to BODC

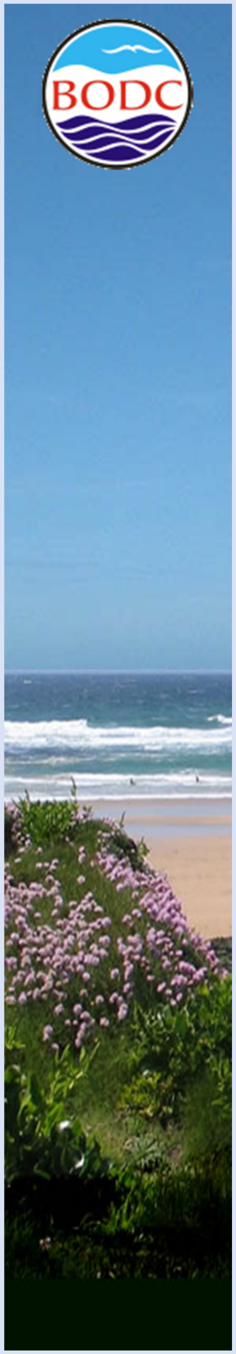
- ➢ BODC Published Data Library

  - Quick-fix solution for IODE/SCOR/WHOIMBL Library and NERC SIS projects
  - Dataset created as a physical file export from BODC data system into a web-exposed data vault
  - Stored with reports as a frozen snapshot
  - Files linked to hand-rolled landing page
  - DOI minted and linked to landing page
  - Automated landing page generation from an Oracle back office is work in progress
  - BUT THIS DOES NOT SCALE!



Bringing Data Publication to BODC

- **Request Publication**
  - Just a twinkle in the eye
- **Accession Publication Service**
  - Under consideration
- **Publication-based workflows and versioned file management**
  - Work on specification and formal project management (PID, project board) just started
- **Discovery dataset specification**
  - Work in progress



Bringing Data Publication to BODC

- Pretty good idea what we need to do
- Many staff years of resources are required
- Competition with many other development projects
- Confident we’ll get there
- I just wouldn’t like to say when…….