Collections to Archives

OceanObs ’09
Venice, 21-25 Sep, 2009

Bob Keeley – Canada
Contributors:
Sylvie Pouliquen – France
Margarita Conkright Gregg – U.S.A.
Scott Woodruff – U.S.A.
Greg Reed – Australia
Many of the CWPs

OceanObs’09

Overview

Sylvie Pouliquen has set the stage by describing what has transpired in the last decade.

My paper will treat data flows from collectors to archives.

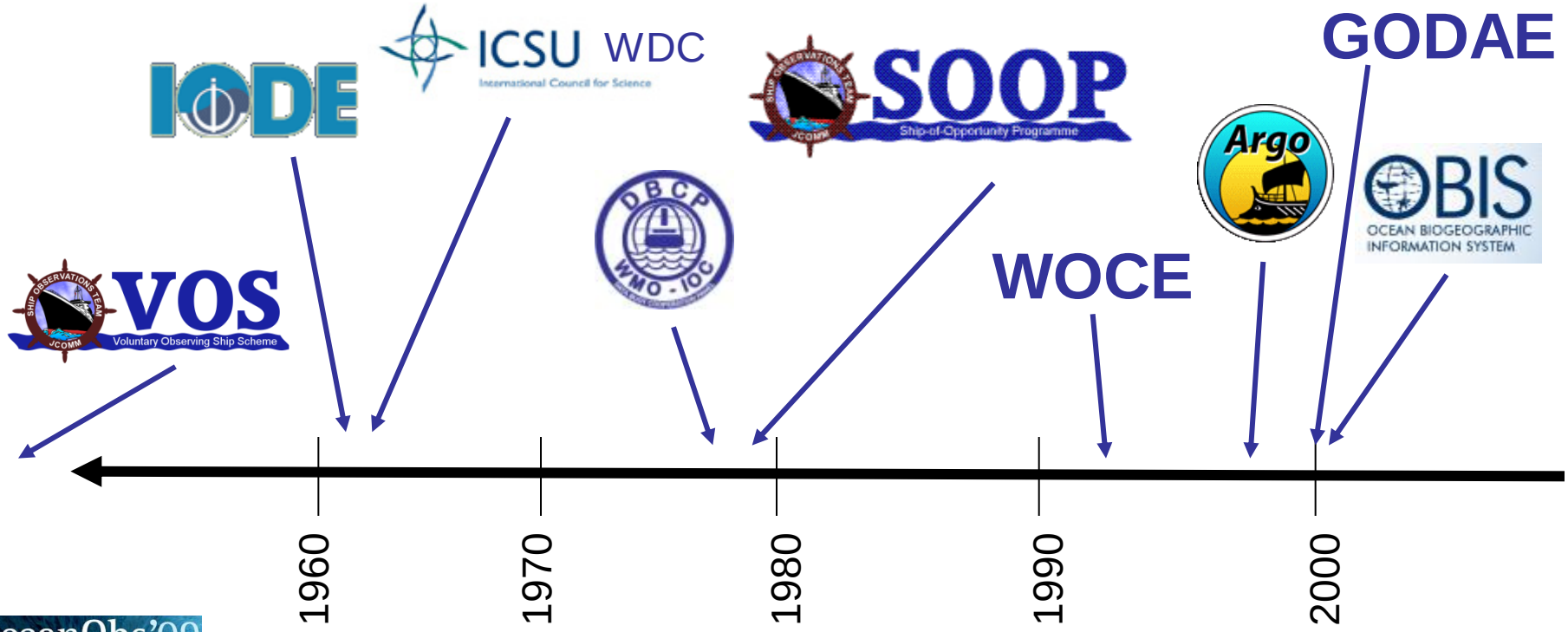
Jon Blower will cover data flows from archives to users.

Steve Hankin will tie all these together with a perspective for the future.

Overview

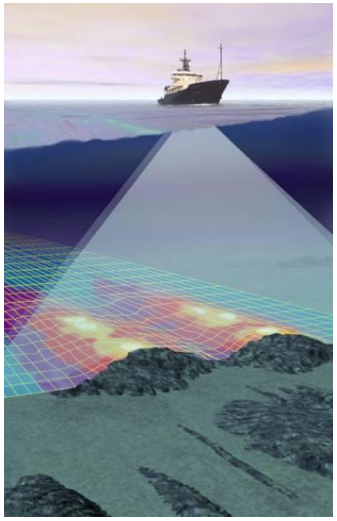
Present data systems are a product of when they were built, who oversaw the construction and what kind of data are handled. This gives us wide variety, many worthwhile ideas and little commonality.

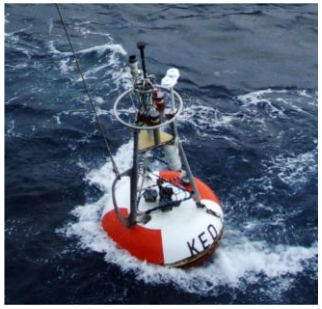
We continue to build systems in isolation, relying on individuals to carry the better ideas forward into new developments.



Challenges

- Increasing data volumes and diversity.
- Breadth of expertise needed to manage this variety cannot all exist in one place.
- Changing technology of instruments underscores the importance of preserving information about the measurement methods.
- Many agencies and individuals are involved in data collection. Knowing them all is hard.
- There is increasing importance of real-time data access.



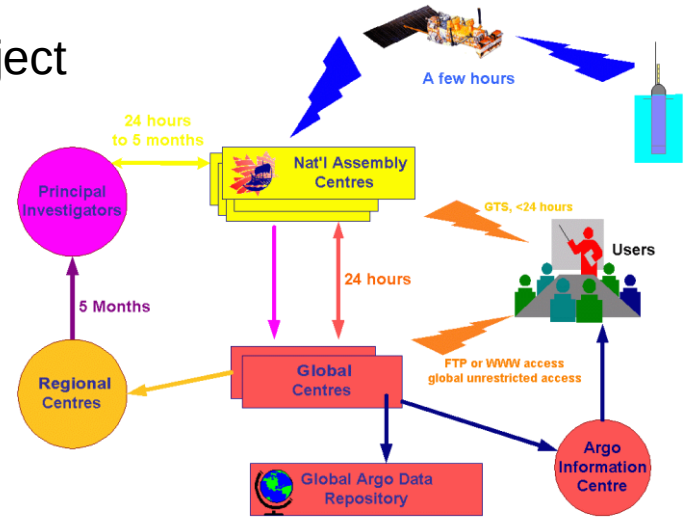


Why Data Assembly

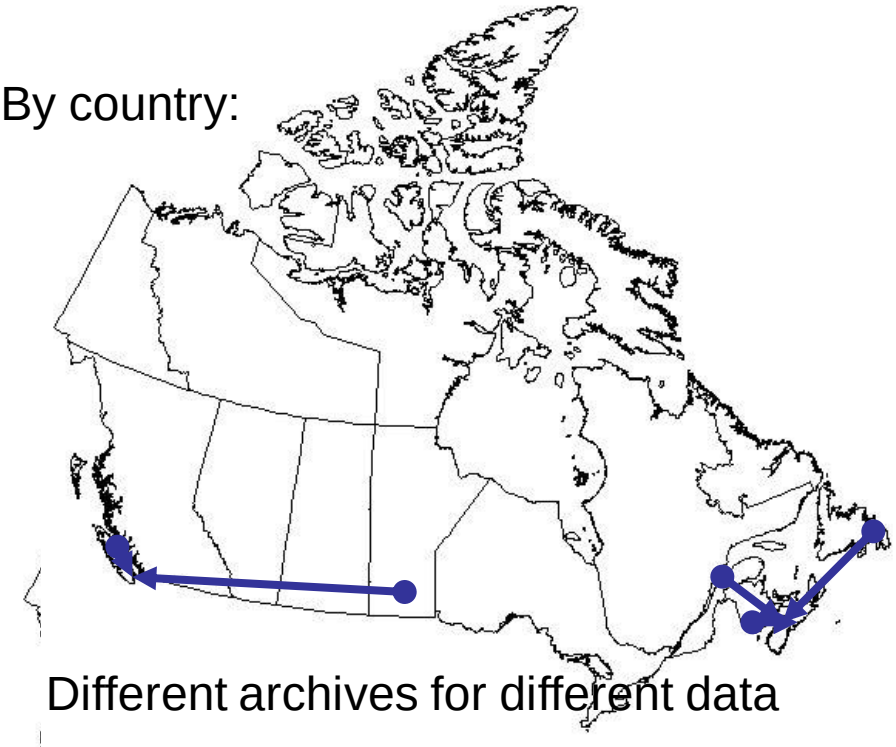
- Uniformity of data structures to make easier integration.
- Uniformity to data quality assessment.
- Fewer data sources to hunt down.
- Standardization of terminology.
- Added value by data merging and consistency of processing.
- Provision of documentation and associated metadata.
- Reduces data management burden on collectors.
- Increases likelihood of preservation into the future.

Types of Data Assembly

By project



By country:



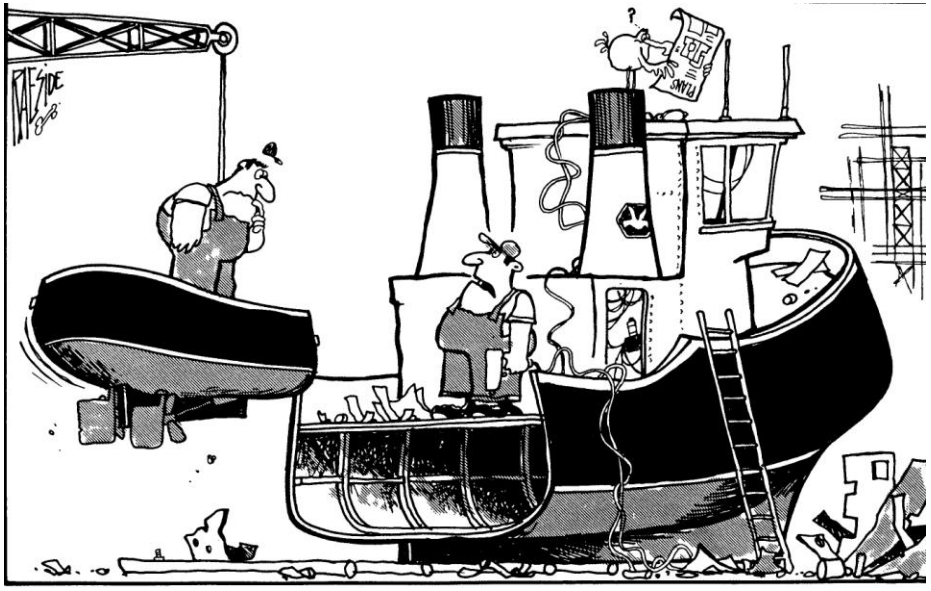
Different archives for different data

OceanObs'09

Sharing

- View of importance varies with type of collector.
- Researchers want to protect IP.
- Private industry wants to protect commercial investment.
- Governments mandate sharing principles but these are not uniformly enforced.
- A number of studies and articles lately dealing in the importance of data management and sharing:
  - "How do your data grow", commentary in Nature V455 4 Sep 2008
  - "Policy Making for Research Data in Repositories: A Guide", DISC UK Datashare project, May 2009, http://www.disc-uk.org/docs/guide.pdf
  - "Motivating Online Publication of Data", M. Costello, Bioscience V0l 59 No 5, May 2009
  - "Ensuring the Integrity, Accessibility, and Stewardship of Research Data in the Digital Age", Committee on Ensuring the Utility and Integrity of Research Data in a Digital Age; National Academy of Sciences, 2009.

Standards



Was that feet or meters?

There are many areas where standards will improve interoperability.

- Vocabularies (e.g. variables, taxa, instrument names).
- Discovery metadata.
- Processing (e.g. quality control, browse features).
- Metadata content (e.g. provenance, instrumentation, methods).
- …….

Assembly targets

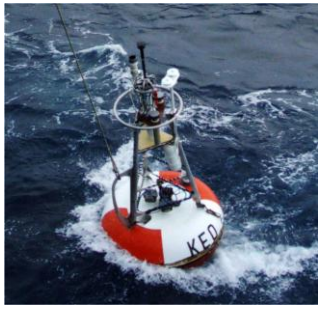
1. Early and close cooperation between archives and data collectors to plan the transfer of data to long term archives.

2. Data centres contribute, assess, recommend, adopt and implement standards as quickly as possible.

3. Improve data exchange formats so that they are more capable of handling the variety and volumes of data.

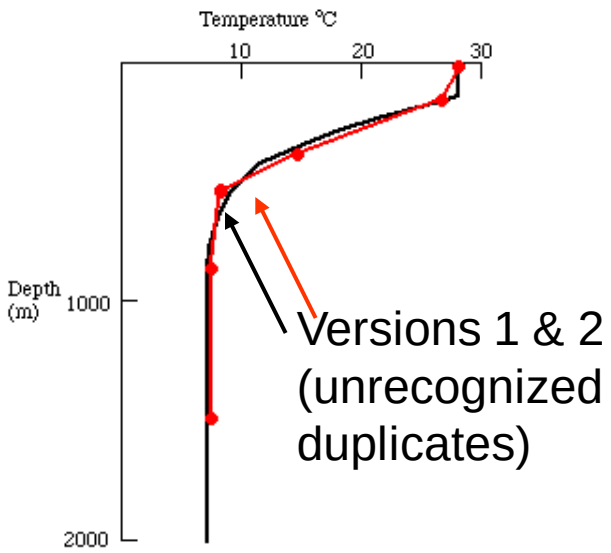
Processing and QC

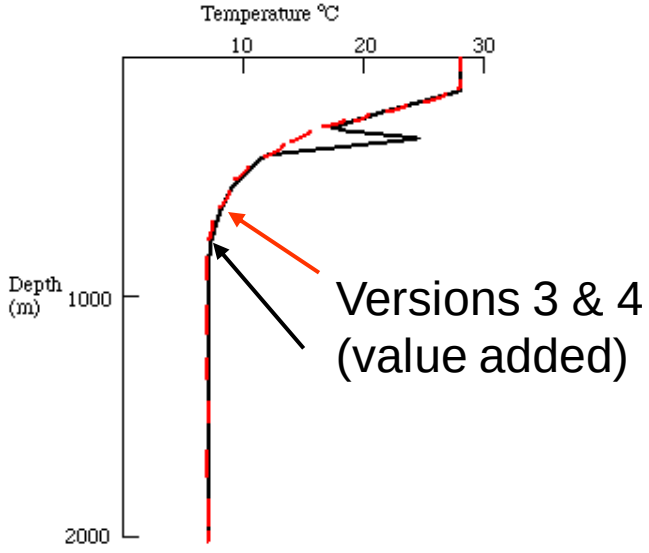
- A variable is measured by a variety of instrumentation, with differing precision, accuracy and methods.
- The variable should undergo common QC, with testing influenced by consideration of how the variable was measured.
- QC by experts should augment that done by data centres.
- There needs to be standards for indicating reliability of the measured value for intercomparability of observations.
- Original values must be preserved regardless of whether any changes are made.
- Clear and easily found documentation of the procedures is needed.



Duplicates and Versions

1. Duplicates (or near-duplicates) are copies of the same data that arise because of limitations or mistakes in transmission, processing, or other activities.
2. Versions may be unrecognized duplicates or value added.
3. Duplicates are undesirable, but versions are to be expected.
4. Duplicates can be detected through unique tagging.
5. Versions need documentation of provenance and content.





Documentation

1. Record history of processing to allow separation of versions, change control, problem isolation.

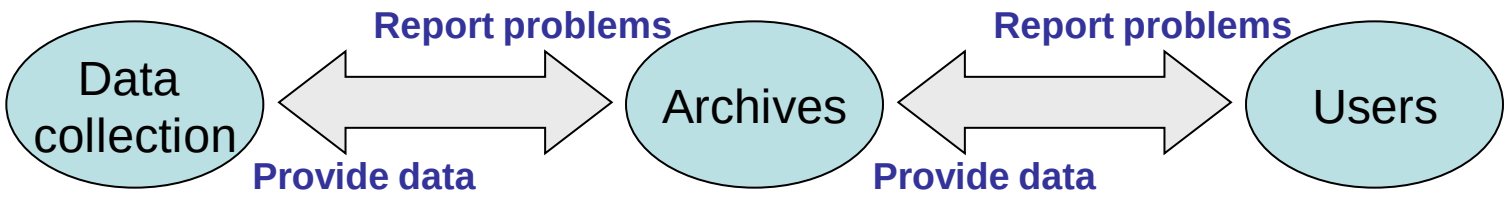
2. Record qc flags, document their meaning and the tests.

3. Include references for algorithms, controlled vocabularies, etc.

4. Describe data sets and products through standard metadata.

Improving archives

1. Chain of processing requires R-T access with quick QC followed later by more careful scrutiny.
2. Some problems will escape detection and will reach archives.
3. Need community who use the data to identify and report problems back to archives to be fixed/flagged. It is highly desirable to have research community contribute to expert QC.
4. Changes in archives emphasizes the importance of version control, documentation.



Data are gathered and undergo quality checking by collectors

Archives carry out quality checking and disseminate data

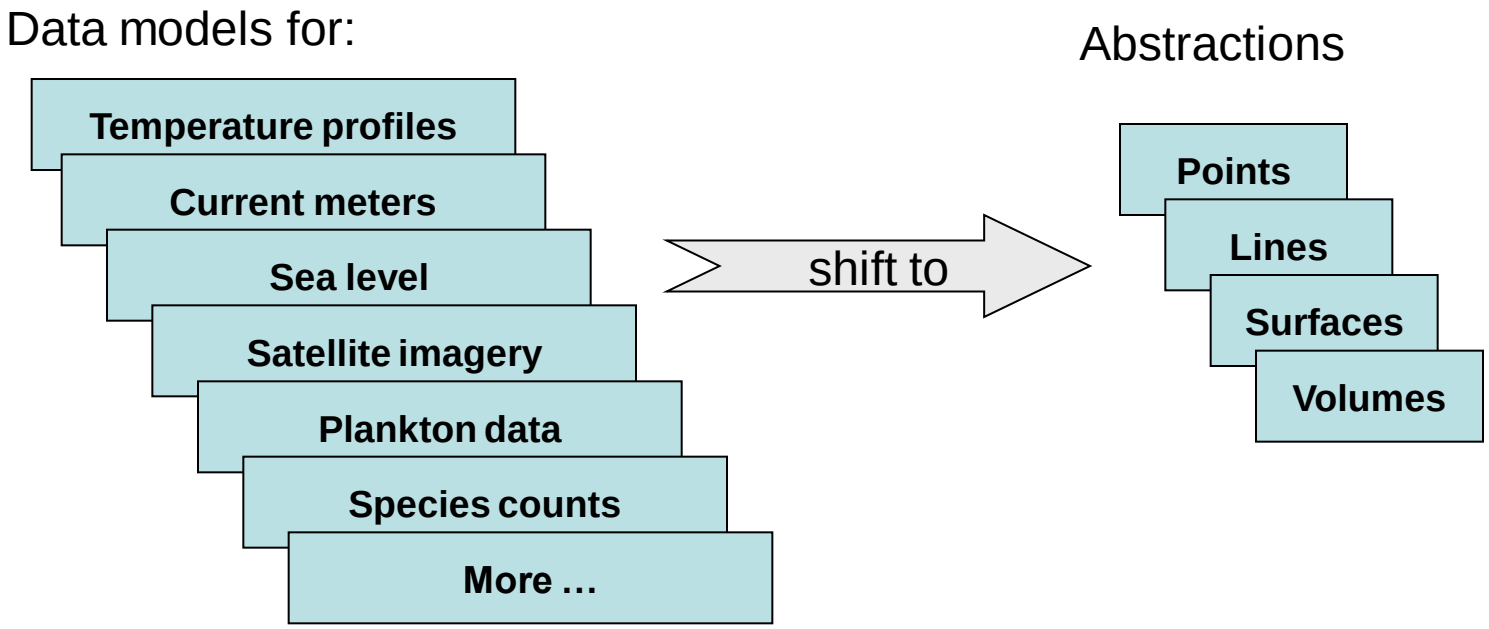
Users will detect problems in the provided data. Reporting these back to the archive improves data collections for others

Processing targets

1. Unique data tags ( or A23F67CC15)
2. Keep processing history.
3. Improved and readily available documentation on qc testing, etc.
4. Better communications between data providers, archives and users when problems are detected

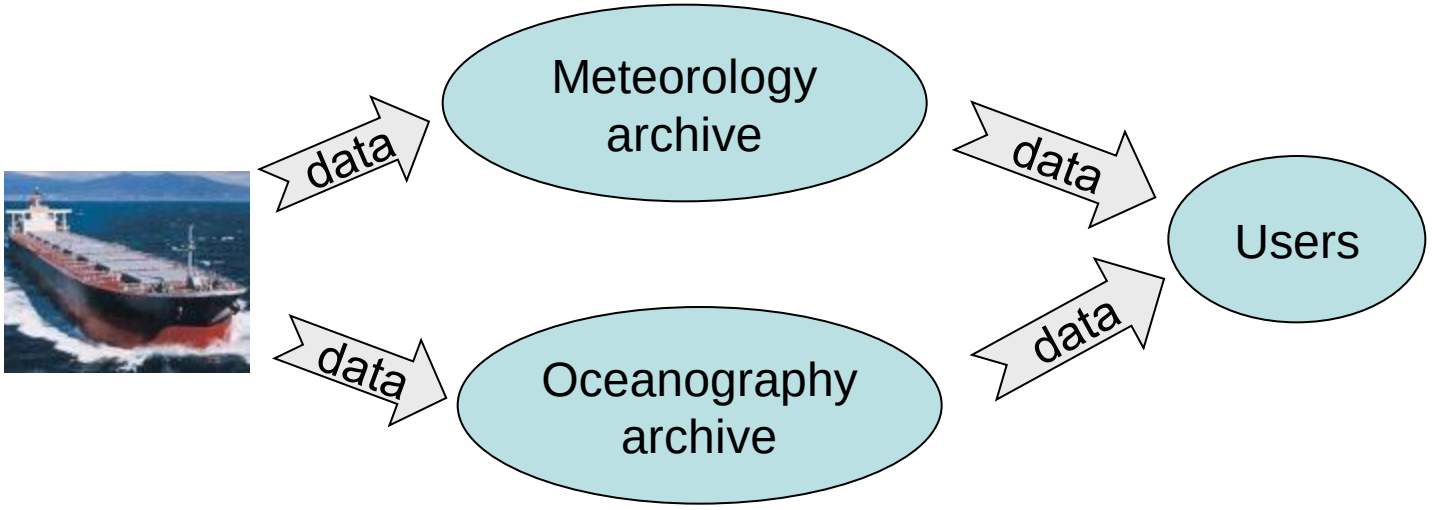
Archiving

1. Most present archives are built with many different data models.
2. This makes adjustments to new variables and new sampling, more difficult.
3. Need to generalize data models to make them more robust to change and utilize standards as much as possible.
4. This should reduce the number of models and help improve interoperability.



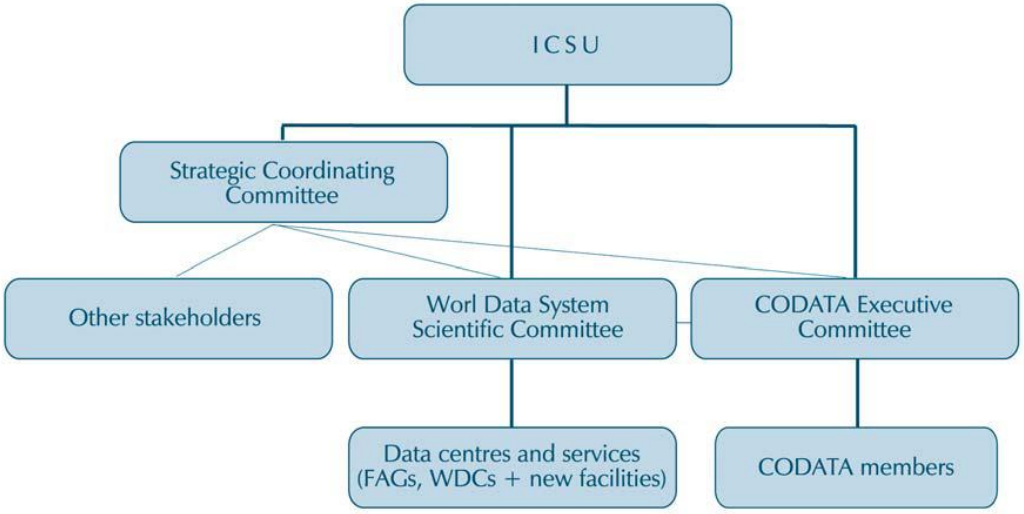
Share work

1. Expertise/familiarity with data is a strong asset for managing data.
2. Increasing diversity of data puts demands on expertise that cannot be met at one place.
3. We must share responsibilities possibly based on types of data
4. We must develop a distributed system that allows data from the same collection program but in different archives to be easily reassembled.



WDS role

1. WDS model is more varied than the original WDC definitions.
2. These may fulfill role of sharing data assembly by type of data.
3. These can focus on product delivery in cooperation with an assembly centre.
4. Offer secure storage into future of data types with no assembly centre.



Proposed new ICSU structures

http://www.icsu.org/Gestion/img/ICSU_DOC_DOWNLOAD/2123_DD_FILE_SCID_Report.pdf

Archive targets

1. Upgrade data models to be more generic/robust to change.

2. Expand and share expertise on kinds of data between archives.

3. Build partnerships with reseachers, other archive centres, and WDSs to ensure all data that are collected have a “home”.

Timely Delivery

1. Rapid delivery of RT forces an assembly line or operationalization of processing.
2. The delivery mechanism is more likely to be push than pull, scheduled than opportunistic, more standards oriented, and likely will use different infrastructure for dissemination.
3. But users must be able to find the data.
4. Let Jon speak to the rest.

What we have to build on

1. Network of data centres in IODE (http://www.iode.org/)
2. JCOMM/IODE standards process (http://www.oceandatastandards.org/)
3. Work done by projects such as SeaDataNet (http://www.seadatanet.org/), IOOS (http://ioos.gov/), Australia (http://www.aodc.gov.au/)
4. Work done by OBIS (http://www.iobis.org/), GBIF (http://www.gbif.org/), taxonomic groups (http://www.sp2000.org/)
5. WIGOS (http://www.wmo.int/pages/prog/www/wigos/) and ODP (http://www.oceandataportal.org/) developments.
6. Experience of recent projects (http://www.jcomm.info/) such as GODAE and Argo.
7. IOC and JCOMM data strategies.

We are missing the overview that shows where all these pieces fit together and the work that must be done by national “volunteers”.

But we do have the building blocks to begin.

The way forward

1. Convene a meeting of data system developers and maintainers
   a. from remote sensing and different disciplines of the in-situ oceanographic community.
   b. to discuss strategies employed, lessons learned, and to seek common solutions or common developments needed.
   c. follow on meetings will be needed to address specific components.
   d. begin under the auspices of the JCOMM.

2. All projects must contain a data management component
   a. to address how the data resulting from the project will be managed and migrated to long term archives and to users.
   b. developed jointly with the archive and funded at the 5-10% level.

3. National administrators, data managers and journal editors must find a solution to provide career enhancing recognition to researchers who provide data to publicly available archives.

The way forward

4. Data managers must make use of the IODE/JCOMM Standards Process
   a. to submit suggested standards
   b. participate in the assessment of their suitability
   c. implement recommended ones in a timely way
   d. this must be monitored by IODE and JCOMM

5. IODE must encourage data centres and monitor progress towards addressing the many technical details that appear in the plenary paper.

6. Representatives of ocean data systems (data centres, IODE, JCOMM) must have a formal seat in the ICSU WDS governing structures to be better connected to the evolving WDS.

7. IODE and JCOMM must provide a well publicized reference site for data management information, standards, etc. There are the beginnings of this in the JCOMM Catalogue of Best Practices and this must be expanded.