

An aerial photograph of the ocean showing a series of parallel waves moving from the top left towards the bottom right. The water is a deep blue color, and the wave crests are white with some spray. The perspective is from a high angle, looking down at the water.

Ocean Information Technology Pilot Project

1. Telemetry and Communications

Four study groups

- Availability and capability
 - Bandwidth, 2-way comms, connectivity, ...
- Requirements for climate, open ocean
- Requirements for coastal regions
- Polar / remote regions (?)

2. Standards and Protocols

- At least five separate activities
 - a) XML/SGML Project
 - Start *now*; build on existing activities; ...
 - b) Metadata standards
 - c) Non-physical / unconventional
 - Assess the state-of-the-art, identify issues, ...
 - d) “Archaeology”
 - Reprocessing of existing original data into modern standard
 - Use several Pilot studies to scope problem
 - » CalCOFI? CPR records? ISOS? IO exp? Selected historical cruises
 - » Will be tied/linked to other themes
 - e) formats

3. Datum and Data Set Integrity

- Develop methodology for uniquely identifying original data and variants
 - Argo is being used as one test-bed
 - Original data given *and always retain* tag
 - Variations referenced against original
- Study of requirements and methods used in other businesses
 - In situ, satellite, unconventional data, ...
- Scope reprocessing / archaeology
- Connectivity to other disciplines
 - E.G., Meteorology

Some essentials

- Unique Identifier (Tag)
 - remove the problem of being able to recognize several forms of the same data
 - Argo has addressed the issue
 - subsequent processing needs to preserve these tags.
 - make the tag creation process functionally independent.
- Original Data
 - Original data should always be available
 - all measured variables will not all be available at the same time
 - Method 1: Tag is used as the link between different versions
 - Method 2: All version have a complete history (eg, GTSPP)
 - Not all processes will be completed at the origin

4. Data Circulation and Service

- Architecture for data serving and exchange
 - Largely internet based
 - Requires success in themes 1-3
 - Similar to that contemplated for meteorology
 - Dave Fulker presentation
 - Data packets and sets arrive with standard metadata and “tag”
 - Systems like IDD are used to “push” data to routine users
 - The sources may be real-time or high-quality delayed-mode
 - All “global” data are circulated among the GISCs
 - But there will be regional subsets with restricted distribution
 - Each ocean data distribution centre has instruction sets for distribution; event driven
 - The GISCs would retain data for some period
 - There would be “climate” centres who would
 - receive one copy of all (relevant) original data
 - one copy of all “scientific quality” data sets
 - the “tags” would enable duplicate identifications
- Also request-reply (pull) servers
- Make more use of IT compression techniques

New Functionality

- The XML standard (or an equivalent) ensures all data are properly described / characterized
- The needed metacode can be generated automatically
- The “tags” allow identification and removal of duplicates
- There is immediate identification of non-conforming data insertion
- No need to be a major centre to participate (provider or user)
- Traditional data archive (backup) could be automated
- Centres might specialise in a type of service
 - Routine versus ad hoc / itinerant
 - Sophisticated versus non-specialist
 - Strong link to User Interface theme

5. Product Exchange and Service

- Similar arrangement to data service
 - a) Study of available technologies
 - NVOADS project has effectively started process
 - b) Study of requirements (broad)
 - c) Establish regional and/or specialised prototype
 - d) ...
- Need to “permit” innovative data base methods, e.G. MARS server at ECMWF
- User driven (see theme 7)
- “Tuned” for efficiency and effectiveness
- ...

6. Data assembly, quality control

- More formality, accepted procedure
 - a) Study of requirements
 - E.g., climate change, SoE, industry, operating standards
 - b) Study of existing practices for QC
 - c) Study of existing practices for assembly
 - d) ...
- Accreditation
 - FGGE / CEOS Levels 0, 1, 2, ...
 - At level 2: a) no QC; b) auto; c) scientific
 - Institute A, A+, A++ system, esp. for c)
 - Recognize value adding
 - Recognize scientific involvement
 - Agreed peer-review system
 - ...

JCOMM DMPA Jun, 2002

Produced 4 actions

- Study common protocols for SOT area
- Examine data serving technology
- Explore xml for data sharing
- Study of relevant telecommunications and computer technologies.

Brussels Meeting - Nov, 2002

Produced 3 actions

- Establish a group to model metadata. Group to be decided by Jun, 2003. First meeting by late 2003.
- Link to the US DMACS group and to the WMO group on the GTS
- Establish a group to deal with data assembly, QC. Group members decided, Neville as chair.

Standards & protocols / New functionality

Part of this was to examine what SGXML was doing
Meeting held in April, 2003

- Pursue metadata reference model and coding
- Pursue xml bricks generalization to other data forms

Data Set Integrity

The situation:

- Data coming from platforms at sea are at lower than instrument resolution

AND/OR

- when data are distributed in real-time they often have “unusual” data removed.

The result:

- Matching real-time to original data must consider these differences.



Argo operations - 1

The Plan

- All data will report on the GTS within 24 hours
- All data will go to GDACs within 24 hours

Reality

- Most data go to GTS within 24 hours
- GTS is the sole means of distribution for some real-time data

Unique tagging of data at GDAC

- Each float has a unique identifier, never to be re-used
- Each profile receives a cycle number and this is in the data stream
- So the unique tag for a profile is Float number + Cycle number



Argo operations - 2

Data Stream

- Data reported from float is maximum resolution available
- Data reported to GTS has “bad” data removed, 2 decimal precision
- Data going to GDACs has all data reported (original) and 3 decimal precision..

So

- If data go to GDAC, we do not need to match real-time and original data
- If real-time data only go to GTS that is all we will ever see
- Original data will all go to GDAC
- Most desirable data are original data



Ship operations - 1

The Plan

- All data will go on the GTS as soon as possible
- All data will go to archive centres as soon as possible

Reality

- Some data go to GTS in 30 days or less
- After 10 years, O(50)% of real-time data still only version in archives
- Because of delays in original data arriving at archives, they can look substantially different from the real-time data.



Ship operations - 2

Unique tagging?

- The ship identifier is unique to a ship
- There is nothing attached to a profile to make it unique since times and positions may be altered after better navigation is attached to the data.

So

- Implement a CRC32 tag



Ship operations - 3



"Best" data now

Full resolution
on disk

SEAS process

Create BATHY
Calculate CRC

BATHY

GTS

Full resolution
with CRC

Calculate CRC
and add to data

MEDS

BATHY with CRC

NODC

Match CRC of full
resolution to BATHY



Ship operations - 4

Future

- This solution can be applied to other data providers independently
- It has application even if/when BUFR is the distribution format of data on the GTS
- It can apply to other situations with adaptations



Data Circulation and Service

US DMAC - Ed knows more than me

