DATA MANAGEMENT BEST PRACTICES

Committee Meeting - 3:00pm Tuesday, December 11, 2007
Parc 55 Hotel, Tuscany Room, San Francisco
Attending: Arko, Chandler, Clark (guest), Graybeal, Johnson, Miller, Prince

Discussion Summary

The first face-to-face meeting of the Committee resulted in a lively discussion of the agenda topics:

1. Review candidates for UNOLS standard cruise-level metadata fields.
2. Discuss the identification of a set of UNOLS recommended standard data products for a cruise
3. Discuss opportunities for funding UNOLS metadata and data preparation for archives
4. Discuss methods for maintaining open communication with vessel operators, techs and scientists

The meeting provided guidance for the development of a cruise-level metadata schema, which will be created by Robert Arko for review by the Committee. Additional time will be required to refine the list of UNOLS standard cruise data products, and to identify funding opportunities. Further technical information is available online at http://data.unols.org/meetings/2007-12-11-agu/20071211_AGU_topics.pdf. The next meeting of the Committee is planned for the March 2008 Ocean Sciences meeting in Orlando, at a date and time to be determined.

As evidenced at several recent meetings, there is a significant gap between operators' and scientists' perspectives on "who is responsible" for data management on UNOLS platforms. A comprehensive solution will require a major cultural shift in the community as well as significant new funding.

This Committee is not charged with laying out a comprehensive solution. We are only charged with
1. Identifying existing best practices, and
2. Making selective recommendations for new ones.

Our work is an "important next step" toward building community consensus for a wider effort.

Our work is guided by these two broad principles:
1. The community is best served if metadata are *routinely reported* by *every platform* to a *central repository* in a *standard format*.
2. It's easier to train a dozen operators than a thousand scientists.

We recommend this initial set of information to be routinely reported for each cruise leg:
1. Basic "cruise-level" metadata record
2. Navigation (platform track i.e. time+position)
3. Event log ("everything over the side")

Much of the basic cruise-level metadata is *already* routinely collected by the operators and/or the UNOLS Office, and needs only to be standardized and reported to a central repository. Every ship in the fleet today navigates by GPS and could report a simple lat/lon file to a central repository. The UNOLS Office can act as the central repository in the near term for prototype work.

Every cruise must be assigned a unique and persistent identifier, including transits that are charged days.

This identifier is the primary key for metadata, publications, etc. The central repository should arbitrate (verify uniqueness) and/or assign new identifiers as needed.

Our proposed "cruise-level" metadata schema will follow these design principles:

1. each element is marked *required* or *optional*, and *repeatable* where appropriate.
2. we will adopt existing controlled vocabularies where possible, but this may be difficult in sticky cases like "device type." The central repository should accommodate aliases, i.e. alternate spellings of ship names.
3. the schema should be sufficiently generic to extend beyond UNOLS someday.

Navigation can be considered to be metadata. Barring exceptional circumstances such as classified missions, navigation should be made routinely and immediately available, not placed on proprietary hold. Once platforms have reported their navigation to a central repository, further shoreside processing can be done such as:

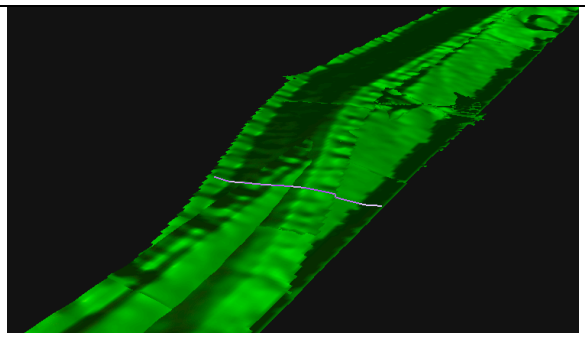
1. reformatting, resampling, editing, annotating;
2. calculating control points (abstracted trackline) and bounding boxes;
3. evaluating quality and comparing systems.

Every ship with satellite connectivity could ideally transmit metadata directly to a central repository in near-realtime. Examples of services already operating include:

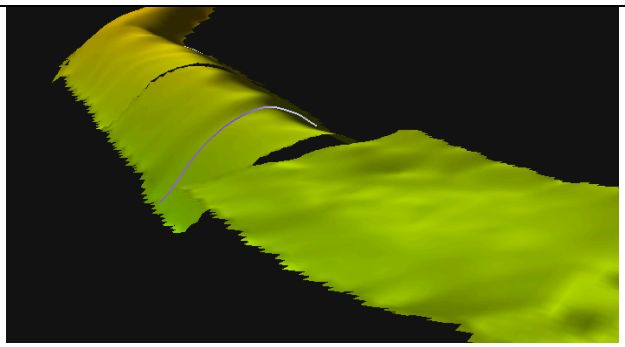
- Since March 2007, *Thompson* routinely uploads daily DAS/IMET files (navigation + met/ocean data) via ftp over HiSeasNet to the MGDS central repository
- Since May 2004, *Healy* routinely transmits event logs, aloftcon camera photos, multibeam statistics, and daily data reports via email over CGDN to MGDS.

There appears to be a solid consensus that the current practice of relying on chief scientists to disseminate data is problematic, especially with standard underway geophysical data and multibeam. While a broad community depends on the routine flow of shipboard data to add to our understanding of worldwide seafloor structures, shipboard biologists, chemists and physical oceanographers are focusing on their own science programs, as they should be, and usually do not have the time or training to make sure multibeam data are appropriate for dissemination. Sound velocity and roll bias corrections require a systematic approach. It appears that many scientists and project and national repositories are unaware of the quality issues of shipboard data products disseminated by chief scientists in recent years, and assume that modern data quality is as good as in prior eras in which institutional processing was funded. (Examples of typical data problems have been inserted below.)

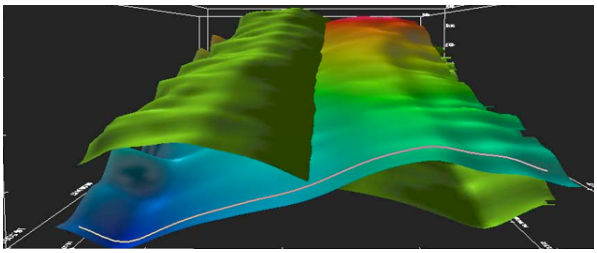
Bathymetry requires attention before releasing for scientific use



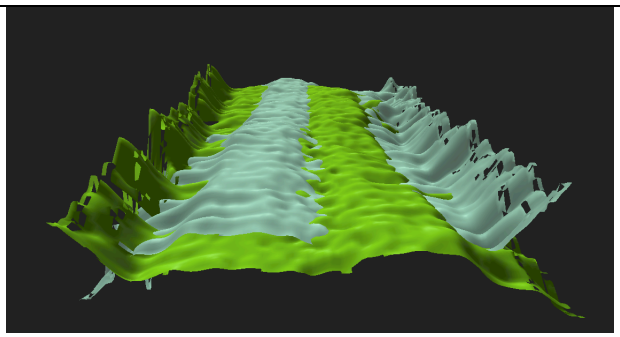
Maps made from shipboard versions of data reveal systematic artifacts.



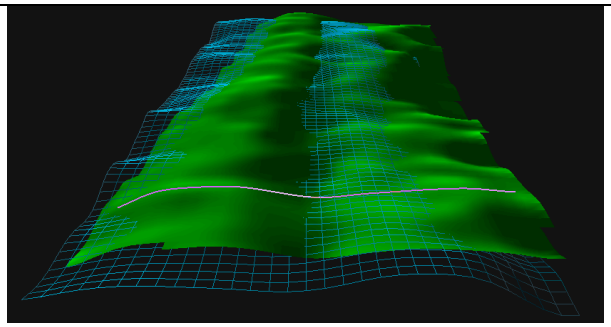
Correct sound velocity needs to be applied to all swath files.



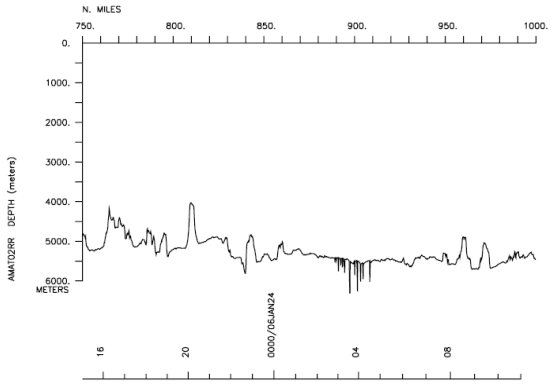
Best shipboard version of roll bias indicates need for recalibration.



Artifacts appear at outer edges of swath, even after roll bias corrected.



Test visualization of two tracks with artifacts removed and roll-bias correction nearing perfection.



Artifacts in center beam depth need to be removed before underway merged geophysical data can be released.

There is a growing awareness that ship-operating institutions need to play a stronger, central role in quality control and dissemination. At the same time we hear the loud and clear concerns of the shipboard technicians, as they are mandated to take on an ever-increasing number of responsibilities, without increased staffing. There is a need for improved tools, procedures, and of course support to make shipboard data quality control more efficient, and more effective.

It is critical that information related to over-the-side sampling events be captured on board the ship. Curators and scientists have complained that inaccurate and conflicting after-the-fact data entries for time, lat, lon and sample identities have hampered the re-use of the samples. Shipboard technicians have complained that they are so busy with other tasks that they can't be "note-takers for scientists." A more streamlined and automated approach to sample tracking needs to be achieved, reducing the amount of manual data entry. Event logs should be as comprehensive as possible (not merely "the station work"), and timestamps must indicate time zone. (A proposal was submitted by SIO and LDEO to streamline sampling for rocks and cores, and automate the flow of metadata to repositories. However we have learned in January 2008 that it was declined. Proposals for improvements in infrastructure face challenges in competing with new field programs, given the current limited resources in the science programs at NSF/OCE.)