

Bob Vrijenhoek, Dave Clague
Feb 8, 2005

1. There is an area we have not covered in our opportunities which deals with discretely sampled data. We have been focused on observatory and "streaming" type data management, but have not talked about data management of samples (which might be framegrabs, physical samples, gene sequences, etc.)
2. Science would like some visualization tool that can take very disparate data sets (format) and put them in to a 3/4D visualization with bathymetry included (like layers). Currently most are using GIS stuff (ArcVIEW) to do this, but it is by no means easy. Mike M's 3D vis stuff is great, but it takes work to get the worlds built and they are limited as to how much you can put in there before it collapses on itself. Another interesting aspect would be to have the bathymetry in a real-time dive display. TerraServer was mentioned here. (Kevin's note: TerraServer really only serves images and could be used to serve up the bathymetry, but you would need something like WorldWind to serve the visualization needs and integrate the data ... assuming it is easy to get external data sets into WorldWind). Also would like actual dive video included in these visualization (not just framegrabs). "Scale of data" was mentioned a bit here. It was along the lines that we can't view data of largely different scales in the same application. The main thrust of this is that scientists want a way to tie ALL observations together easily so they can look for new and interesting correlations (Kevin' note: this one sort of has a 'mission-statement-ish' feel to it). Basically all of the data we deal with has temporal and geospatial aspect and we should be able to visualize things that way. Of course, modeling data would also have to be integrated.
3. The key to get science use it to make it easy and fast. NGDC has a way of entering data, but there is a huge time lag.
4. One very important key to getting users to enter data is to make it easy for them to enter it while it is being collected. Otherwise it will never get entered.
5. We should be curating our samples like museums do and maybe have a dedicated person doing this. This would include some sort of labeling (bar code system) (Kevin's note: Rich Schramm has been wanting to do this with PDA and barcode readers for sometime ... he would be an excellent contact). This was born out of a concern that when scientists go away, so do their samples so we should be curating them for future use. Also, if we curate them properly and get metadata somewhere, it will serve as a cost savings for others who don't have to repeat collection (i.e. Dave's finding rocks in Scripps dredge logs).
6. MBARI is unique in that, per scientist, there is WAY more data than anywhere else (Kevin's note: looks like that should be in our strengths).
7. Quality control of data is another huge issue. Even if we have an easy way for people to enter data, we must have QC mechanism to flag/identify bad data. Don't get rid of it! Just flag it! There was talk about peer voting on data quality (like Amazon's customer review). When people enter data, there should be some server mechanism that gives them an automated first cut at how their data looks. If the QC is bad, people won't use the data system. MBSystem was talked about as a good example here.

8. Data security and embargos is a MUST. The big concern is that if you have people put data in repositories without proper embargos, data miners can quickly use the data to publish results.

9. Data should be integrated not collocated. The network is becoming powerful enough that data should be distributed and not have to be stored in one repository. However, there must be data security/integrity standards that all of these distributed servers use so as to never lose the data. (Kevin's note: Dave Caress alluded to there being lots of ways to do this out there, but "they would take enormous amounts of effort". We may want to ask him more about what these are to look at competitive/collaborative efforts).

10. A great number of data sets from instruments have good management systems in place (gene sequencing, seismic data, etc.), but acoustics is one that is not really well managed. (Kevin's note: should this be in opportunities?). Talk to Stephen Dziak (sp?) at PMEL (Oregon) about this.

Some science questions I heard:

1. I want to compare data from this dive site to this other dive site and they are very far away (GIS can't do this well).
2. Show me all samples along a dive track (like what Mike's 3D vis does).
3. I would like to see where all organisms with this gene sequence were collected in area X?