Controlled-Vocabulary Working Group Report – LTER IM 2006

The controlled-vocabulary working group conducted a session including all LTER IM's and representatives from NCEAS and ILTER at the 2006 Colorado Information Manager's Meeting. The working session included a brief description of the activities of the working group since the 2005 meeting. These are summarized in a Powerpoint presentation at: http://gce-lter.marsci.uga.edu/lter_im/2006/app/uploads/workgroup1/Controlled_VocabularyIM06.doc and in the Spring 2006 Databits (http://intranet.lternet.edu/archives/documents/Newsletters/DataBits/06spring/#4fa). These efforts had been aimed not at developing a new controlled vocabulary, thesaurus or ontology, but on providing resources to help evaluate the utility in the LTER context of existing controlled vocabularies, thesauri and ontologies such as GCMD, NBII, GEMET and WORDNET resources. The planned aim of the working group was to discuss next steps in terms of the evaluation of these resources and discussion of tools for exploiting them to aid in LTER data discovery, and two separate groups focusing on these issues were to be established. However, following the presentations, the IM group decided that a fuller reconsideration was needed, so four concurrent working groups simultaneously addressed the same three questions:

1. What do we want to do? Or more specifically: What capabilities do we want to provide for data discovery and data integration?
2. What existing resources could we use and what can they provide?
3. What do we need to do to develop our own resources?

These working groups then reported back to the entire group. Overall conclusions drawn from those presentations were that:

1. Groups were complementary – each group focused on different aspects of the problem
2. The effort was worthwhile. A high priority was placed on improving the ability of LTER data to be searched and browsed.
3. The challenge is complex. It was suggested that the best route was to start simple, and look at other efforts and how they are addressing the challenge.
4. There was a discussion of how many words are useful. A vocabulary with too few words leads to lack of precision while too many lead to a lack of reliability.
5. We need to involve working scientists in the process, since they will ultimately be the "consumers" of the system.
6. We need to know which search terms researchers are actually using to locate data.
7. Information managers need to be involved in the evaluation of alternatives.
8. The products and tools will evolve – this is not a one shot thing.
9. We need to be aware of the substantial work meeting this challenge will require.
10. We need to have an appropriate balance between top down vs. bottom up.
11. We need to decide how broad (in a disciplinary sense) we need to make our vocabulary.

Based on that discussion, we reached 4 major conclusions/action items:

1. We should enable auditing on Metacat to track requests so that we can see how researchers are attempting to use the existing system.
2. We need additional educational activities to help prepare the LTER IM group to deal with the complexities of thesauri and ontologies.
3. We should continue efforts to examine what others have done and to relate that work to our own efforts, such as taking a closer look at NBII, GCMD etc. are doing/have done
4. Continue existing work on compiling lists of attributes and work with the work with SEEK KR group to represent attributes in ontology template.

Appended below are raw notes from individuals participating in the workshop.

NOTES FROM DUANE COSTA: What should we address?

1. What do we want to do? What capabilities do we want to provide for? What tools, how to apply it?
    - browse engine
    - tools to make new suggestions, formal mechanism for nomination?
        - a tiered system, where a term gains recognition at the site level and is nominated at higher levels formally
    - Should the vocabulary describe what people are working on, or aim to be broader and allow browsing to find specific things hierarchically?
    - Controlled vocabulary could be a **keyword source** for journal articles
    - The controlled vocab could be used as a **synonym finder** for documents that already have keywords. Better than trying to mark up old documents.
    - We need a tool to **mark-up** existing datasets with new keywords.
    - **Thesaurus** to make suggestions about preferred keywords to use (based on an EML document or full text; find all terms that are already listed and suggest new ones to be accepted or rejected.)
    - If every dataset had EML with attribute descriptions, do we need more tagging to build up browse interface? People are generally looking for data, not new concepts. We would just have to set up a **higher level ontology** to help organize it. This would not generally work for publications.
        - According to SEEK, every observation is an entity (stream, stem, leaf) and one property (length, diameter, flow) on it.
        - If every attribute had a keyword representing the entity, every term would return a dataset. Builds effective searches.
        - Can we identify the ‘entity’ and create a hierarchy to organize them?
        - How does one classify a timestamp like this? Is someone likely to search based on a time? Of what entity is time a property of, and event?
        - The next step along this path would be to divide existing keywords into entity/property, and use entities to seed the keyword list.

- Do next: Set up a mechanism to grab measurement attribute data, names. Low-hanging fruit? The problem is that the description doesn't always give the entity – it may be "flow" instead of "stream flow", or "sample" instead of "oxygen". Requires human input.

2. What existing resource could we use and what do they provide? Existing thesauri?
   - For a given term in a controlled vocabulary, which is preferred (CO2 / carbon dioxide)?
   - What are synonyms?
   - Parent term/child term hierarchies exist in some thesauri
   - How do we use and apply them?
   - Is there a vocabulary of junk words somewhere we could use to winnow?

3. What do we need to do to develop our own resources?
   - We've already looked at existing documents, EML, bibliographic titles. We've made lists of keywords and definitions, over 20,000. Started narrowing down. We have a list of 1000 words we think are solid; may want to grow it.
   - How large are comparable vocabularies? 3000 – 5000?
   - Should we focus more on terms dealing with measurements?
   - Do we want to include non-English terms? Even in English, vocabulary varies.
   - Possible next step (from above) – grab and **analyze measurement data** to work bottom-up.
   - Determine the shape of the browse tree, and whether to go top-down (may include terms not used much) or bottom-up (requires someone to work out the ontology). Or perhaps both.
   - Create a database of all the search terms, and use it to see what hierarchical level people want to find data at in metacat. The current 1000-word list ranges over a wide spectrum; a "**target level**" would be nice to know. Modify metacat interface to record searches.
   - Start organizing a **hierarchy** from the terms we have.
   - How do we involve the **broader community**?
     - Ask NSF for money to do a workshop.
     - Other external agencies doing similar work, such as NBII. Develop a network of our partners to share ideas.
     - Cuhasi, OpenDAP, cleaner, mmi (marine metadata initiative)
   - Envision the three **top levels** of the tree.
   - Define **who** is browsing so we know who to aim at. Scientists at the LTERs, or IMs.

NOTES FROM MARGRET O'BRIEN:
Background on the Seek project (from MarkS):
People: Computer scientists, @Davis (Bertram, shawn)
Postdoc, @ecoinfo Josh

Searching:
Search for data object (a package would be returned in our case)
Find a column in a table (if you look at the attributes and move one step up, and annotate)
A data entity may not appear to be match your search until you manipulate 2 cols (eg, biomass and area => density)

Define “Observation”: is a repeated measurement of a thing’s characteristic|trait|property

Need a mechanism of annotating a data object

There is an ontology for ecology called Oboe (=Observation Ontology for Ecology)

Sidebar: GO (= gene ontology) a list of 1000’s of terms with very little depth. But the common vocab gives them a framework to talk)
So these (controlled vocabs) have great use, even if superficial

Flicker, “folksonomy”
Wikipedia (synonyms linked)
Thinkweb (fernando villa)

Controlled vocabulary
Glossary (give meaning to the vocab)
Context of vocab