Data Access Working Group (DAWG) Meeting
New England Center, Durham, New Hampshire
May 28, 2004

Meeting Participants: 2004 DAWG Meeting Participants.pdf

The DAWG meeting was opened with a welcome from Bruce Caron, DAWG chair.

The 2004 DAWG meeting focused on a number of issues. These included discussions around:

1) Report and learned from the DLESE Data Services Workshop that had been completed the day before. This discussion was lead by Tamara Ledley
2) Details along the data stream: anastomosing data flows from DAAC to science to education. This discussion was lead by Cathy Manduca
3) Developing the Data Access Pattern Language. This discussion was lead by Bruce Caron
4) Data FUSION and related interoperability issues: getting data into layers that allow multi-layer analysis and visualization. This discussion was lead by Ben Domenico.
5) Planning for the 2005 Meeting

Lessons Learned from the Data Services Workshop – Ledley

At the Data Services workshop participants were divided into teams that were built around data providers which gave the small teams focus. This proved very effective in having the teams successfully begin the development of educational modules in the form of Earth Exploration Toolbook (EET) chapters. The poster session allowed for cross-pollination between the teams. Workshop was successful in having participants enter their tools and data in Using Data Portal (**http://www.dlesecommunity.carleton.edu/usingdata/resources.html**). Educational and Technical reports will be developed from the notes accumulated on the Data Services Workshop swiki (**http://swiki.dlese.org/DataSvcsWkshp-04/1**) along with the completed EET chapters (serc.carleton.edu/eet). It was emphasized that DLESE Data Services (DDS) should put their resources to completing these EET chapters and understanding the path it takes so that we can replicate it.

It was noted that there is a large gap between powerful tools developed for research purposes and scaling back to meet needs of education. Simple interface (menu) to choose features of the IDV that you want/don't want in your tool. You don't have to know JAVA to get what you want. A model for this is WISE (Web-based Inquiry Science Environment, **http://wise.berkeley.edu**), which provides inquiry based activities for grades 4-12. It is not clear the extent to which Earth science datasets are included in these activities

Groups are beginning to request input on the use of data in education from the DLESE

Data Services team. This gives us an opportunity to be leaders in the effort to get more data used in other disciplines as well as ESS. The groups that we have interacted with include:

Landsat Team at Data Services Workshop – Should there be a workshop focused on what scaffolding is needed to make Landsat data useful in education. The formatting of Landsat data and its distributed location of free Landsat scenes is an issue (May-June 2004)
RODES Workshop (July 2004)
Planetary Science Community (March 2004)
International Polar Year – Bridging the Poles Workshop (June 2004)

In light of this the question was asked: How do we take the experience we have with the Data Services Workshop and with these individual groups and scale it up? One problem that was encountered was that some groups who were invited to participate in the Data Services Workshop were either "not ready" or "too focused on their science" to consider working with educators. It was suggested that we write and distribute case studies describing how groups have been successful. This will encourage us to write a really readable (propaganda) piece that describes what we learned and and what they can do: not just a distillation of what happened, but something that highlights the benefits of the learning an implementation of data best practices. This could be used by DLESE Ambassadors. We can point to the EET chapters as a concrete outcome of the meeting that a very broad community can use.

Some suggestions for future DATA Services Workshops include

- Have sessions where end users rotate around to do hands-on work with various tools right along with the tools developer. Like a CRIT lab where you set up a scenario for participants to learn
- Part of the workshop focused on reviewing portals/interfaces. It was suggested that this was to broad and that a criteria along with a rubric be developed before we have participants review the portals/interfaces again.
- It was also suggested that we have the educators come to the workshop with a problem that they want to solve.

At this point the discussion evolved to the issue of **having your data in DLESE** and whether having an EET chapter means that your data is in DLESE. One problem that was perceived is that relationship between DLESE and Using Data Portal as well as between DLESE and EET is not clear. A new model for data in DLESE shouldn't replace the Using Data Portal. We aren't starting from zero. Some groups have already begun integrating data. DLESE shouldn't render obsolete any of the investments that have already been made.

We also need to know how users, contributors, resource creators find the data they need. How can we clarify where people can find data in the form they need it, whether a dataset available for using in educational materials or already wrapped in an educational context.

It was suggested that educators probably do not look for data by data name but by region. It was thought that a spatial search option might be useful.

Details along the Datastream: Anastomosing Data Flows from DAAC to Science to Education – Manduca

During the last meeting it was noted that the pathways for data are manifold and complex with lots of players. It was suggested that having information about how data is processed, ie the lineage, would be useful for curriculum developers. Jim Frew has been working on lineages for a few years. A lineage system might describe the processing process as the data travels through the process. With XML it might now be time to create a lineage community to build systems for tracking lineage.

If we were to track how a dataset has been changed we would also be able to provide a new type of metric for the data provider, and provide another type of metadata.

Tracking lineage is not only useful for the educational community but also for the cross-disciplinary research community. Within discipline information about data processing is known by individuals and is not documented. As a result, in cross-disciplinary research the loss of this information can be very detrimental. With data crossing disciplinary boundaries more often now, we have the opportunity to describe a broad-based lineage system with standard mechanisms for gathering metadata and the ability to query them.

Compound documents may be a way to make the lineage of a datasets visible to the user. Compound documents are created by different groups, each of which adds value to the document. The document would have imbedded in it pointers to datasets. One type of compound document would be an inventory list used to create a catalog.

The THREDDS group is trying o associate data with textual documents. This would allow search technologies to find the information in text that would lead people to data.

Basically there are three things that would help make data more available and usable. These are:

1. Data providers have tools to build catalogs of different classes of data. Long lists of pointers to simple lists of data and how to access it would be useful to curriculum developers. Curriculum developers would find the description in DLESE and then go to the list to pick the data they need.
2. We should also facilitate connections to resources that provide access to specialized datasets that focus on specific scientific issues, ie El Nino, SST, and are wrapped in rich educational materials
3. Cataloging facilities should have the extended capability to include lineage metadata.

<u>Using Data Portal</u>
The site contains data sets and tools. It has been upgraded to include a defined range of classes of data to search on. Users can find things by “Ease of Use”, “Resource Type:

Dataset and Tools", or "Data Type." Within the category "Resource Type: Datasets and Tools" the defined options currently are:

- Images, maps, animations
- Datasets
- Tools
- Datasets with Tools
- Datasets with Teaching Activities

Within the category "Data Type" the defined options currently are

- Observational
- Real time
- Model/Simulated Data

These categorizations need to be examined by a broader community and extended/refined.

Once the metadata for data is determined it will be necessary to develop a mapping of the metadata into a schema that the user would use to find the dataset or tool they need. Using alias terms for metadata will help to lead people to the sources. "Word NET" is another way to map associations from definitions to natural uses.

Issue: "A Place for Data in DLESE"

The issue was brought up during the last DAWG meeting. The question is can there be a place where raw materials can be exposed to product developers, which are marked a not being end products?

We need to take the initiative on providing feedback for the quality effort.

The current situation is that we don't have catalog inventories in DLSE, but current metadata can show observational data, in-situ data, or remotely sensed data. There is an inconsistency in cataloging (if the developer doesn't do it) where the "data" metadata box is not checked. It may also be that activities may just dump a user on the front page of a data providers site without any information on how to search for or get the data.

There are cataloging issues around having datasets as resources for educational materials in DLESE. What would the metadata be that would effectively catalog these datasets and tools? Focus groups indicate that keywords were the best way for searching. However, to make this effective those developing the metadata need to think like developers.

Another questions that was raised was should there be access to incomplete educational materials and the datasets they contain? Maybe call the whole thing the "Developers Site"

Perhaps the Using Data Portal could be used as a prototype of the "Developer Site" Funds are available through DLESE Community Services to make a demonstration (experimentation) site. Those interested in helping to make a recommendation on this are Bruce, Katy, Tamara

Concept Mapping

As an experiment to see what the lineage of some data resource are the DAWG broke into smaller groups and did some concept mapping. The instructions were to

1. Choose a single raw dataset and put it in a circle
2. Draw a line from that circle to indicat a person ding some processing to the data
3. Circle any product that contains the processed data

There were 5 concept maps created during the DAWG meeting. The datasets represented on these concept maps included, DAAC, Landsat, MODIS, NOAA Data, Weather Data. Digital images of these concept maps appear at the bottom of the swiki page at http://swiki.dlese.org/dawg/22.

Types of Data Products

- Instrument read out (for instrument team)
- Unprocessed data (for processing groups)
- Processed quality data (for scientists, educational product developers)
- Research data products (combined datasets, interpreted datasets (for scientists)
- Images of data
- Processed image products e.g. animated GIF's
- Educational activities
- Data products for educators
- Research papers
- Papers for the public
- Informational sites
- Data products for industry/commercial use
- Numerical forecasts

Key Players in Using Data

- People who work with the data
- Instrument teams
- Scientists
- Data Massagers
- Educational Teams (curriculum developer/scientist/technical)
- Faculty
- Tool developers
- People who run and finance data operations
    - Managers
- End Users
    - Teachers
    - Industry
    - Public/policy makers
    - Students
    - Curriculum Developers
- Others
    - Catalogers/Archivist/ontologists
    - Access providers to collections

- Censors/ethicists/philosophers

As an outgrowth of this conversations, that at the previous DAWG meeting, and from the strands at the 2003 DLESE Annual meeting, the DAWG decided to develop a formal recommendation concerning "Having a Place in DLESE for Data" During the meeting a draft recommendation was formulated and over the next six weeks it was refined by email. While the Steering Committee was not ready to deal with the issue they were made aware of the recommendation. The final version of this recommendation follows.

Recommendation from the Data Access Working Group to DLESE
[finalized June 25, 2004]

"To facilitate DLESE as a place where participants with varying expertise can work together to create finished products that will ultimately become part of the collection, DLESE should provide a clearly-identified repository for production materials – including data, metadata, draft educational modules as well as analysis and visualization tools that are under development.

To avoid frustration on the part of DLESE users seeking finished, reviewed educational products, this production environment should be clearly identified as such.

The DAWG will move forward on this by implementing a prototype of such a system on the "Using Data in the Classroom" portal, <http://serc.carleton.edu/usingdata/index.html>."

Developing the Data Access Pattern Language – Caron

The Earth Data Access Pattern Language Project would be designed to assemble the available knowledge about the data access production and use cycle (focused on education) by constructing a set of patterns that illustrate the relationships between the technologies, the science, software application development, and education.

The process of creating this set of patterns begins with discussions among data providers, technology builders, and education data users that reveal their needs and problems, as well as lessons learned (good and bad).

These discussions are distilled into the patterns, that is, into descriptions of various circumstances along the data access process, and insights into the problematics and solution fields that are integral to these circumstances.

Examples of how particular solutions (e.g. best or worst practices) have proved valuable (or horrible) illustrate and articulate the solution field.

The goal is to have a dynamic digital document that is of real value to data providers (e.g. the DAACS), technology builders (like THREDDS), software application builders (like

the Data Discovery Toolkit and Foundry), scientists (like ESIP IIs) and teachers and students.

One of the benefits of doing a pattern language approach is that it offers a learning opportunity to its builders... it is a revelatory process that mines tacit knowledge and returns shared knowledge. As such, it is the cornerstone for a community of practice among earth data users.

The process will begin in a more formal way at the next DAWG meeting with a half-day devoted to patterns.

Between now and then, the challenge to the DAWG members is to begin to assemble their lessons learned and to look at some of the pattern language resources.

It might be useful to get initial input through this volunteer process which might show enough potential value to get some modest funding (e.g. for a workshop) to do something more substantial.

Some background information on Patterned Language is available on the swiki at **http://swiki.dlese.org/dawg/24.**

Data Fusion and Interoperability Issues – Domenico

One of the most confounding and fundamental problems with data is the issue of comparing disparate data, i.e. data in different file formats such as GIS files and gridded data. This problem is a barrier to people who are trying to do cross-disciplinary science – at the forefront of Earth system science. Studying floods, for example, requires a lot of different data in a lot of different forms including stream locations, weather forecasts, locations and phone numbers of schools in potentially affected areas.

How might these disparate types of data be fused to make them more useful in a cross-disciplinary way. Are there sufficient example of fusible datasets that could be brought together with curriculum developers to make good materials?

An example tutorial was shown about remotely sensed data, preprocessing of data, and data fusion. It included theoretical background as well as small examples of fusing data. It also includes information on data compression and case studies. The module is copyrighted and is not currently available for distribution.

We should consider something like this as a case study for development and demonstration. Perhaps we should enlist the support from some other group. Identifying and displaying good data fusion examples might be a better use of the DAWG resources.

What should the scope of the DAWG be?

- Promotion of data tools?
- Content creation?

- Promoting standards for data?

There should be a registry of tools and formats?
DODS can do much of this transformation behind the scenes for you
Plan to move forward on the Develop's site.

Next Steps:
- Keep adding things to the Using Data Portal and refining the metadata structure
- Check Developer's Workshop Report for specific requests
- Perhaps this site becomes a collection
- Items that it might contain
    - Authoring Environment/Tools development area
    - Highlight work in Photoshop as data preparation
    - Collection of vector graphics that could be re-used
    - Tutorials on this items
    - THREDDS materials

Planning for the 2005 DAWG Meeting – Caron

The DAWG is in an advisory role to DLESE Data Services and through them to the Management Council and Steering Committee.

Work that is currently being done that contributes to data
- Earth Exploration Toolbook (EET)
- Using Data in the Classroom
- Using Data Portal
- ESIP Federation
- REASoN
- GEON
- NSDL
- DDTF
- THREDDS

We should contact those people who wrote letters of support for DLESE Data Services and get a description of what they are doing…what they are contributing.

Next DAWG Meeting:
Dates: 1.5 days during March 2005
Place: Santa Barbara

Agenda Items to Consider:
- Focus on data discovery
- Use a patterned language approach to come up with patterns that articulate the data discovery process
    - Who would be the beneficiaries of this exercise?

    - What product would we product that the community could use?
    - Possible idea:
        - Produce a draft outline of the patterned language
        - Present at the Annual Meeting
        - Based on feedback move forward of give it up.
- Ask would (DLESE groups) want our input
- Ask ourselves what expertise we bring to this effort that would be valuable to DLESE
- Consider the four priorities from the last DAWG Meeting

Other Activities

1. Support/contribute to the development on an experimental/demonstration site on the Using Data Portal. Produce a draft of standards to consider: Cathy, Tamara, Kate, LuAnn, Katy. This experimental site will have a section for "incomplete" materials – undocumented datasets and tools that are still under development. We need to consider how DLESE can be a leader in this area. Should eventually talk to NASA's Data Access Working Group like people.
2. Ben will facilitate the effort to finalize the draft statement on creating a place within DLESE for datasets and tools.
    a. Need for standard data access protocols that is adhered to by both data suppliers and analysis and visualization tool developers
    b. A reporitory for raw, unpolished material, including data, metadata, and modules and analysis and visualization tools under development.