EarthCube White Paper

Meeting Scientific Requirements: How Much Can be Done Locally?

Karen Remick(kremick@iarc.uaf.edu) and James Long(jlong@iarc.uaf.edu)
International Arctic Research Center – University of Alaska Fairbanks

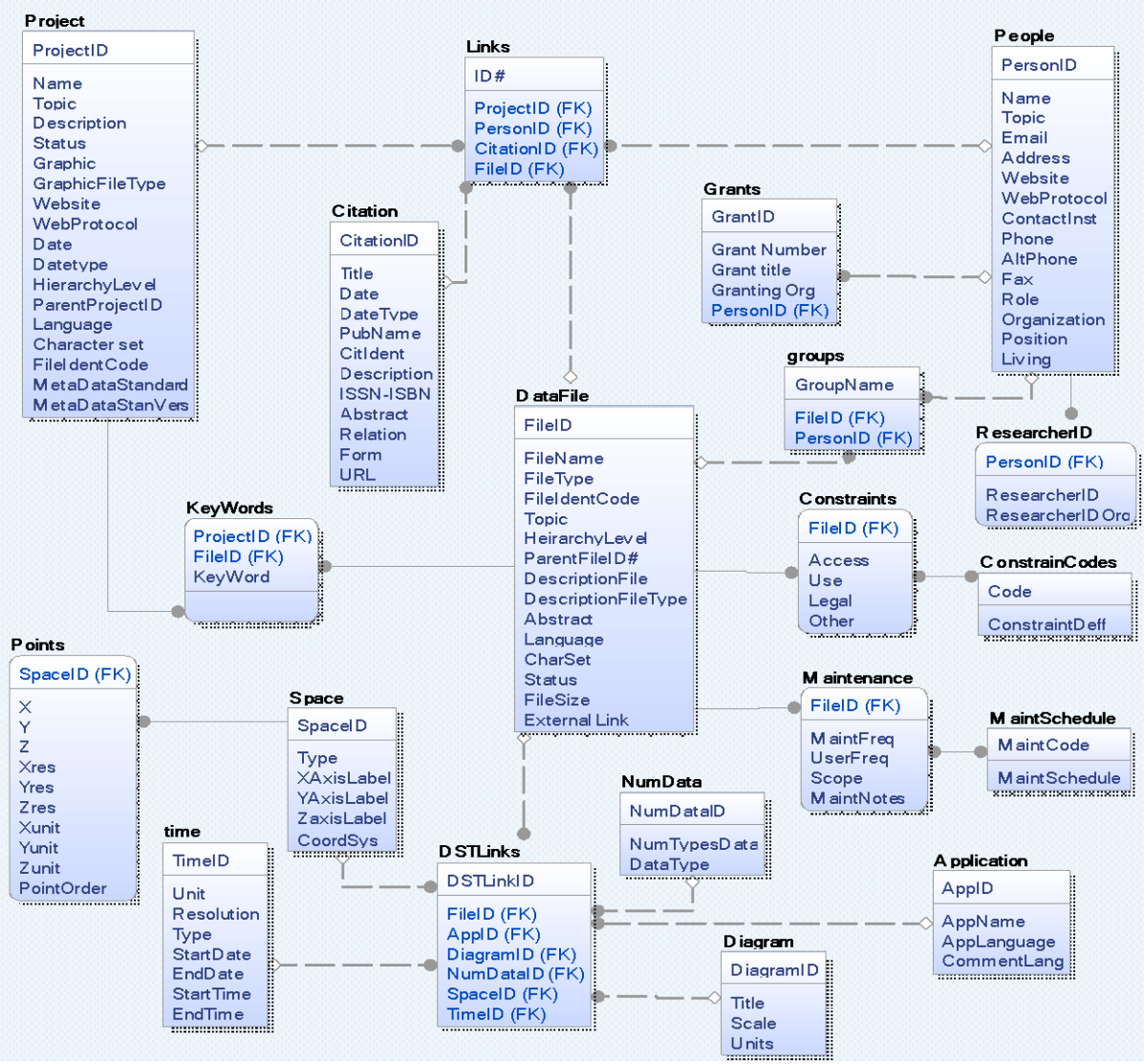
Introduction

There are many anecdotal stories about a researcher finding the key piece of information he needed for inspiration or a breakthrough in an unexpected place. While not data, they serve to illustrate the importance of cross disciplinary studies. Thus data centers must be able to support data from many disciplines in order to support a given scientific community. Another important service for data centers to offer is professional networking. Once a researcher decides he needs data from another discipline, he may need an expert in that field to help him determine what data to use and how to use it appropriately, or he may need new data taken. Development of networking tools for scientists has lagged behind social networking and other professional networking, as scientists need both networking tools and data, and current websites either offer a list of data with no networking tools or networking with no data (Lackes *et al* [2009]). While EarthCube seeks to correct this issue, much can be done at the level of individual data facilities to improve networking abilities and facilitate cross disciplinary studies. The key to effective scientific networking is organizing the metadata such that relationships are defined. Authors linked to papers liked to studies, etc… Is the associated paper a report of results or a description of the model used? This white paper discusses the way the International Arctic Research Center at the University of Alaska Fairbanks proposes to organize their metadata in order to enhance data use and networking. Looking at this system we can see what is possible for a single institution, and what can be done when the search tools are replaced by those of a larger organization.

Conceptual schema

In order for data of all types (numerical, blue prints, models, etc…) to be accepted into the system, the structure of the database must mirror the structure of research rather than a characteristic of the data. Organizing data by project is not a new idea. At the Science and Technology Facilities Center in England Mathews *et al* [2010] use the CSMD model, a project based organization model that they have been using and refining with good results for the last 8 years. Projects can have multiple parts and each part can have multiple data sets, multiple people working on it, and multiple publications associated with it. A hierarchy needs to be definable when needed, and the system needs to be searchable by project name, topic, time, space, working group, people or citations associated, data set/diagram/application name or key words. Below is the current schematic for the future IARC database. It is still in development and proposed names and constructive criticism is welcomed. The scheme is organized around two wheels: the ***Project***, ***Datafile***, ***Citation*** and ***People*** tables form one wheel which defines the project, and the ***Datafile***, ***Numdata***, ***Application***, ***Diagram***, ***Space*** and ***Time*** tables form the second wheel which defines the data. The ***Links*** and ***DSTLinks*** tables establish the relationships, the spokes of the wheels. This system is designed for maximum flexibility in data. The data accepted can be numbers (the data table) pictures or maps (the diagram table) or computer models/applications (the application table). A given project can contain any combination of data, diagrams and applications, for example a

model may have a numerical source data set as ***Numdata***, the model itself as an ***Application*** and ***Numdata*** and a ***Diagram*** as the result. The ***DataFile*** table contains an entry 'description' that allows for the detailed description of the experiment, model instructions, assumptions made and/or anything else that would help a future researcher understand and interpret the data. The ***NumData***, ***Application*** and ***Diagram*** tables can be associated with the ***Time*** and ***Space*** tables when applicable. The ***Space*** table allows for the definition of different coordinate systems. The space type can be point(s), path, area or volume. This enables more accurate representation of the data. The data from a ship that sailed around the coast of Alaska is represented as a path, not a square that includs the state's interior.



The extreme flexibility in search parameters allows users to interact more with the data. A researcher who is skeptical about a paper's conclusions can look up the paper or authors and find the associated project. From there they can find and download the relevant data and description file and check the result for themselves. A student looking for a project can call up a semantic engine that will find all data sets with similar times, locations and/or topics and do a semantic analysis of the abstracts to determine

complimentary data. The student can then get the email address of the PIs with authority over that data so they can contact them if necessary. A researcher who wants to do cross disciplinary work can search the people table under topic they need a collaborator in, review the papers of the resulting choices, determine who's work fits best with what they need and contact the person. With the support of a larger federation or organization, chat boxes, discussion boards, and rating systems can be programmed into the web interface and used to increase the ability of the data center to function as a professional networking center. The following is an example given in both our other paper as well as on the EarthCube website. In this paper we will use it to examine what aspects follow from the flexibility of the metadata organization both with and without a semantic engine, and what benefits a national level organization would bring.

Somewhere in the near future…

Dr. Davis flies above a thunderstorm cloud in New Mexico. The instruments on his plane include a high speed camera, spectroscope, magnetometer and GPS. In Brazil his co-I, Dr. Fernando, flies above another storm with the same equipment. Using a web connection, both researchers are uploading their data in real time to the archive in Fairbanks, AK, and downloading each other's data so that they can see both on the same screen. In addition, their GPS and magnetometer data are being downloaded by Dr. Rogers in Fairbanks. He is running his magnetic field model using the real time data and guiding the two pilots stay at conjugate field points. The archive's web programs allow chatting between all the researchers both in text and voice. Dr. Rogers was suggested to Dr. Davis by the archive's semantic engine as a possible collaborator for this project. Dr. Davis had checked him out looking at the ratings of his publications and his contributions to the discussion boards on the archive's networking sites.

With the storm finishing up, Dr. Fernando brings up the archives analytic tool box and does some preliminary analysis on the data. She sees that the frequencies of certain emissions are higher than in the last storm and starts making notes about other differences she observed between the storms (saw more blue jets this time, etc…) to possibly explain the difference. She has made arrangements to have a copy of the data transferred from the Alaskan archive to the one run by her own establishment INPE.

Dr. Davis, whose last grant proposal had not been funded, had worried about his graduate student whose project it was to be, however his student contacted the archive and asked its semantic engine for some complimentary data sets in geomagnetic fields and it had turned up several options. His student had picked a geological survey with an overlapping geomagnetic survey. Using those and Dr. Roger's model he was working to see if the mineral content underground could be predicted form the deviations between actual and theoretical values in the magnetic field, a project some researchers at the mine has shown interest in.

Dr Davis, flying back to base, is thinking that the definition of the word 'archive' has changed almost as much as the definition of the word 'telephone' when he is interrupted by a notice from the archive. It has done its daily scan for articles that fit his entered parameters and found one that may be of interest for him. It also informs him that Dr. Wilkins received funding for her thunderstorm prediction project. He'll have to get in touch with her when he gets back. He is pleased that the flights went well and that

the data is both safe (already being backed up on the archive's drives) and secure (his security settings allow no-one but his work group to have access to the data until his publication comes out).

In this scenario, the archive has/will provided:

- Collaboration - finding a selection of appropriate people,
    - Locally without a semantic engine –the researcher could search the ***People*** table for the needed Topic, and then use the names to look up the projects they worked on using the ***Projects*** table, and papers they wrote using the ***Citations*** table. They could then determine who should be contacted.
    - Locally with a semantic engine - a semantic engine would search the database in a similar manner to the process above, saving the researcher time and effort.
    - Nationally with a semantic engine – all qualified collaborators nationwide and their qualifications would be supplied to researcher.
- Communication between researchers on the team,
    - Locally – the ***People*** table gives email and website information, so emailing during the operation is a possibility or they could use an external program such as skype or an instant messenger program available on the *Links* page of the archive's website.
    - Nationally – an organization could sponsor a mash-up page using pre-written communication (chat boxes, voice, etc) programs.
- An analytical tool box to allow preliminary analysis of archive stored data,
    - Locally – The *Links* page would have links to free downloadable tools such as the statistics program r, and make available any application written by a member.
    - Nationally – much larger selection of onsite tools usable on a larger selection of data.
- Information about, and easy access to, public data sets,
    - Locally without semantic engine – researcher searches the ***Projects*** table under topic and reads the abstracts. Also links to other archives or public data sets not in the archive could be listed on the *Links* page.
    - Locally with semantic engine – engine reads time, place, topic, and abstracts from the ***DataFile*** table as well as for known data sets on the web, and searches for sets that match the researchers preset conditions
    - Nationally – same but the engine has access to a nationwide pool of data
- New research questions through offering a selection of complementary sets,
    - Locally without semantic engine – researcher searches ***Projects*** table for research whose data files (***DataFile*** table)have similar/complementary times (***Time*** table), locations (***Space*** Table) and/or topics, reads the abstracts for projects and hopes for inspiration. A randomized program that features data sets could get lucky and suggest an appropriate data set.
    - Locally with semantic engine - semantic engine searches through the tables mentioned above along with the abstracts and uses preprogrammed assumptions to find data that is complementary to the researchers past work.
    - Nationally – same as above, but with a much larger metadata set, thus increasing chances that something interesting will be found.

- Monitoring of grants and publications
    - Locally without semantic engine – *Links* page has links to the web pages of professional journals, google scholar, and announcement pages to funding agencies.
    - Locally with semantic engine – Semantic engine searches the sites mentioned above and compares the abstracts to preprogrammed criteria, selecting papers and grants appropriate for the researcher and sends a notification when one meets specifications.
- A variety of levels for data security before archiving,
    - Locally - ***Constraints*** table defines to whom a particular data set can be released to. Authorization and authentication handled locally.
    - Nationally – Same control over the data, however authorization and authentication is handled nationally through a single sign on system, allowing the researcher to be cleared on a number of archives at the same time.

Other possible services not mentioned in this scenario:

- Listing of publications and allowing archive members to rate and comment on them,
    - Locally - publications in the ***Citation*** table can be listed and associated with the project (**Projects** table), specific dataset (**DataFile** table) and author (**People** table), but no rating, tagging or commenting. It is, however possible to contact the author.
    - Nationally – National organization provides the program for the interactive mode.
- Sorting of projects by field, but allowing easy access to other fields of study,
    - **Projects** table can be sorted by topic, key word, citation, etc… so knowledge of any one attribute can bring up the appropriate information
- Member profiles and statements of research interests
    - A webpage with static profiles can be provided by the archive with information populated by the **People** table listing Topic, website and contact information.

All of this is possible using today's technical knowledge. Archives without a semantic engine can accomplish similar results, though it takes much more work on the researcher's part and requires a good search system and good metadata organization.

Conclusion

While the national level organization would vastly enhance networking capabilities and the quality of the search results, project based metadata organization allows for considerably richer semantic connections that ultimately lead to more meaningful search results.

References

Lackes, R. Siepermann, M. Frank, E. *Social Networking as an approach to the enhancement of collaboration among scientists* International Journal of Web Based Communities 5, #4 2009

Williams, B., Sufi, S. Flannery, D. Lerusse, L. Griffin, T. Gleaves, M. Kleese, K. *Using a Core Scientific Metadata in Large-Scale Facilities* International Journal of Digital Curation #1, Vol 5, 2010