EarthCube White Paper

A Data-Based Professional Networking System

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

Introduction

Today, scientific cyber-infrastructure is confronted with three primary challenges. The first is that, due to technological advances, scientific studies are generating data at an unprecedented rate. These data sets must be properly organized, stored, and distributed so that they can be verified and built upon. The second is that interdisciplinary studies are becoming more and more important, thus the organization of data and metadata needs to be flexible enough to address the many forms of data organization used by different disciplines. Finally, because there are more researchers and data sets than is knowable by any one person, a professional networking organization needs to be designed to allow researchers to know who is doing what, and how work from other people affects their research plans. This white paper discusses an approach that addresses these challenges.

There are two primary ways of determining what data is accepted into a data center, subject and location. Which type is used depends on the purpose and location of the data facility. If a data center is subject oriented, such as one found at a medical facility, its data organization is tailored be useful to the relevant discipline. However its use is then restricted to researchers in that subject and it makes cross disciplinary studies difficult. If a data center is location based, and only accepts data from a specific institution or geopolitical location, such as at a university, it needs flexibility of organization to allow for cross disciplinary studies, but restricts its members to seeing only locally produced results. The federation of data sites can overcome the boundaries for both organization schemes, allowing institutions to develop naturally in a way that is beneficial to them, yet allows the free exchange of information.

Vision

It is important for data centers not to be focused on just data, but also on ensuring that the data can be used in different ways. Professional networking advances the use of scientific data in many ways, such as putting researchers in touch with potential collaborators, letting researchers know what other people are working on so as not to spend time and money on redundant studies, and inspires new research questions by exposing a researcher to new ideas and people. At this time, however, professional networking is very limited. While 65%-88% of people in business network, only 15% of scientists do (Lackes *et al* [2009]). Data is what allows scientists to function and must be the center of any professional networking system, but one of the reasons scientists do not network is that sites that hope to encourage it tend to provide just a list of data with no networking capabilities (Lackes *et al* [2010]). A data center's assets must include a professional networking capability, with tools that foster making the necessary connections. This is demonstrated in the following scenario describing how a project-based system with semantically-located networking connections could aid scientific research.

Somewhere in the near future…

Dr. Davis flies above a thunderstorm cloud in New Mexico studying sprites. The instruments on his plane include a high speed camera, spectroscope, magnetometer and GPS. In Brazil his co-investigator, Dr. Fernando, flies above another storm with the same equipment. Both researchers are uploading their data in real time to their respective data centers, and simultaneously downloading each other's data so that they can see both sets 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 to stay at conjugate field points. Professional networking web programs allow chatting between all the researchers both in text and voice. A semantic agent suggested Dr. Rogers to Dr. Davis as a possible collaborator for this project. Dr. Davis then checked out his member profile, looking at his publications and their ratings as well as his contributions to the discussion boards on the networking site.

With the storm finishing up, Dr. Fernando brings up a toolbox consisting of distributed analytical modules 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 while the experience is still fresh in her mind. She has made arrangements to have a copy of the data transferred from her own establishment INPE to the Alaskan archive.

Dr. Davis, whose last grant proposal was not funded, had worried about his graduate student whose project it was to be, however his student contacted the archive after a semantic query for some complimentary data sets in geomagnetic fields 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 from the deviations between actual and theoretical values in the magnetic field, a project in which some researchers at the mine have shown interest.

Dr Davis is thinking while flying back to base, 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 professional networking site. 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 clicks a button to send her an email expressing an interest in her project and asks her if she is interested in doing a collaborative project in the future. 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, professional networking with semantic search has facilitated:

- Collaboration of people,
- Listing of publications and allowing archive members to rate and comment on them,
- Availability of member profiles and statements of research interests,
- Authentication and authorization for real time uploading, storage and downloading of raw data,
- Real time and email communication between researchers on the team,
- A distributed analytical tool box to allow preliminary analysis of archive stored data,
- Cooperation between archives with regards to data handling and transfer,
- Information about and easy access to public data sets,
- New research questions through use of the semantic queries,
- Sorting of projects by a variety of parameters allowing easy access to other fields of study,
- Monitoring of grants and publications and notification when one that fits the given parameters has been found,
- A variety of levels for data security at different stages, and
- Interfaces that are transparent and easy to use

Other possible services not mentioned in this scenario:

- Association of data with publications,
- Notices of meetings – both national conferences and local talks
- Public and members only discussion forums, and
- Easy communication with the networking staff so new services can be requested.

Community-Based Governance model

Each organization will find different ways to organize data and metadata, thus a national or international system needs to develop from the bottom up rather than the top down. As a result the most important role of governance will be to enable communication between organizations by providing professional networking services and a semantic roadmap to the collection of various data ecosystems. Because needs will change, and organizations will find new ways of doing things, it is important to have flexible governance. Thus periodic meetings should be scheduled between people at NSF (as mediators) and those in authority at the various data organizations to discuss problems, and come up with solutions and standards. Discussion threads can be maintained to discuss ongoing issues or ones that arise suddenly. Since the reputation of the data center will drive how willing researchers are to use it, it would be in the best interests for an organization to adopt the standards and protocols found to most effective for interaction with a national professional networking and semantic services system, as it would reflect positively on them.

Conceptual CI Architecture

We propose two components for a national infrastructure:

- professional networking services
- semantic mapping services

Networking Services

Similar to how people normally network on social networking sites, scientists and their data need to network as well. But rather than just connections based on who you know, data connections can have manifold types of rich connections based not only on people, but also on partnerships, protocols, parts, formats, services, analytical tools, etc.

Our main contention is that these rich connections between people and data need to be enumerated and made searchable. Graphs of these connections can be described by RDF (Resource Description Framework) statements.

Semantic Mapping Services

It's clear from perusing the various white papers that there is a significant desire for EarthCube to federate disparate systems. Brokering, DataOne, and iRODS all address this issue, and the theme crops up in other papers as well.

From our perspective, EarthCube should be an environment where all of these different currently existing data eco-systems can live together. A challenge exists, however, when these various eco-systems employ differing federation strategies among themselves. The solution, we believe, is for EarthCube to model this complex environment in such a way that its constituent connections can be categorized, searched, and integrated.

We propose a semantic layer whose goal is to eventually federate the differing federation schemes, beginning with mapping. If EarthCube operates as a professional networking site that provides common services to members, then different data centers can be members of different hubs with connections based on the various partners, people, protocols, and/or parts involved, where the 'federation strategy' is a subset of protocols, and 'software system' is a subset of programs.

The generic recipe for establishing a semantic system can be followed in this endeavor: Ontologies are created that encompasses the essential features of what it means to be a partner, person, protocol, and/or program, and a semantic agent is constructed to crawl eco-system hubs, issuing assertions about the connections to each one found. These assertions are then combined with the ontology and an inference engine to create a holistic picture of the whole in an RDF store which can be queried via SPARQL.

Having such a roadmap for the gestalt of eco-systems is the first step in a journey through the data landscape on a quest for relevant information about potential partners and possibilities. Standard queries custom tailored to each researcher could be run periodically to alert the researcher to developing possibilities and information.

Design Process

The design process for our proposed solution is simple; 1) Implement best practices for professional networking services, with the caveat of richer connections for data, and 2) use those connections as fodder for semantic technologies.

Users can be brought together with experts in social networking technologies for a requirements charrette, followed by multiple ontology design charrettes for describing the various eco-systems based on the nature of hub connections. Software designed to work in a distributed manner across participating sites would share authentication and authorization responsibilities, as well as professional networking services.

Operations and Sustainability

An open-source model is preferred for software, governed by a foundation responsible for oversight.

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