Technologies for Integration into EarthCube

Sara Graves, Rahul Ramachandran, Helen Conover
Information Technology and Systems Center
University of Alabama in Huntsville
sgraves@itsc.uah.edu

Technologies used in EarthCube will come from many sources. Some will be developed specifically for EarthCube, some will be enhancements or extensions of current technologies and some may be used in their current state. Based on many years of experience in varied aspects of geoinformatics and cyberinfrastructure, as well as collaborations with researchers in both the geosciences and information technologies, we envision an EarthCube landscape of tools and technologies targeting different communities, but supporting common data formats, metadata schemas, and communication protocols. Several technologies for possible consideration in the EarthCube cyberinfrastructure are described in this paper.

Needed Capability: Data Discovery across multiple data repositories and/or registries, to include semantic searches

EarthCube will bring together heterogeneous data repositories from multiple disciplines in the geosciences. While these data repositories will typically provide specialized services for their targeted user communities, support for common data discovery protocols will be needed. Several of the NASA Earth Science data centers have been working together with the Federation of Earth Science Information Partners (ESIP) to develop federated data discovery technologies. The white paper from the ESIP Discovery Cluster describes the *OpenSearch, Datacasting* and *Service Casting* initiatives in some detail. The ESIP Federation has also fostered development of community ontologies for Earth science data and services. An early semantic search engine, UAH's *Noesis* (Ramachandran et al., 2006), is currently being updated and enhanced for use in the data and service casting arena.

Needed Capability: Data Analysis Tools

EarthCube will require a suite of data analytics and visualization tools to support geosciences research. Tools that assist in knowledge discovery from large data volumes will play a critical role in the success of EarthCube. The *Algorithm Development and Mining toolkit (ADaM*; Graves et al. 2007, Ramachandran et al. 2007, Rushing et al. 2005) was initially developed in the early 1990's with the goal of mining large scientific data sets for geophysical phenomena detection and feature extraction, and has continued to be expanded and improved. Unlike most data mining software, ADaM has been designed for scientific use with image data. ADaM has over 100 components that can be configured to create customized mining processes. The ADaM toolkit includes classification, clustering, and feature selection / reduction techniques as well as a number of utilities that are useful in pattern recognition applications. ADaM also provides a set of image processing modules that are useful for extracting features from images as a precursor to mining or pattern recognition. Each component is provided with a C, C++, or other application programming interface (API), an executable in support of generic scripting tools (e.g. Perl, Python, shell scripts) and SOAP-based web services for integration in distributed service oriented architectures.

GLIDER (Globally Leveraged Integrated Data Explorer for Research; Ramachandran et al. 2009a, 2009b) integrates ADaM with the Interactive Visualizer and Image Classifier for Satellites (IVICS; Berendes et al. 2007) and NASA World Wind 3-D globe. By combining these tools, GLIDER serves as a powerful tool for detailed analysis of satellite imagery and extraction of thematic information from datasets. GLIDER provides three main views for managing and manipulating satellite data imagery, each with a well-defined purpose. The Project Explorer View allows users to manage data files and analysis results organized by project. The Image Analysis View provides an interactive analysis capability on the data in its native sensor view. A full suite of image enhancement algorithms, band math and classification algorithms can be applied within this view. The Earth View provides a GIS capability for GLIDER. Many satellite images can be simultaneously projected and displayed on a 3D globe. In addition, other data sets can be displayed as layers for comparison and analysis with full geographical context.

Needed Capability: Text Analysis Tools
In addition to data, the EarthCube will likely serve as a repository for research articles and technical papers, data documentation, and other textual information. Text analysis tools will help EarthCube users mine through these resources, as well as open scientific journals, online articles and web pages for information and research results relating to geosciences. *Spyglass* (Rushing et al., 2009) is designed to help analysts explore very large collections of unstructured text documents. Spyglass uses a domain ontology to index documents, and provides retrieval and visualization services based on the ontology and the resulting index. The ontology based approach allows analysts to share information and helps to ensure consistency of results. The approach is also scalable and lends itself very well to parallel computation.

Needed Capability: Science Collaboration Environments
Collaboration and communication are keys to successful scientific research. Conferences and publications remain extremely important venues for scientific exchange, but interest in effective on-line collaboration environments continues to grow, especially with the explosion of social media. By providing a forum for sharing preliminary results, digital data, and analysis tools, such environments can foster rapid knowledge sharing within a virtual community. To support this kind of on-line collaboration, UAH/ITSC has developed *Talkoot* (the Finnish word for barn-raising), a software toolkit to provide Earth Science researchers a ready-to-use knowledge management environment and an online platform for collaboration (Ramachandran et al., 2011). Talkoot allows Earth Science researchers a means to systematically gather, tag and share their data, analysis workflows and research notes. This technology infrastructure also allows researchers to control the sphere

of their collaboration: within their group, within their research center, across institutions or in some cases, with the entire community.   We are also active in the Earth Science Collaboratory effort within the ESIP Federation.

Integrating Capabilities: Data Mining Solutions Center, as an example
The Data Mining Solutions Center (DMSC) Portal, built using Talkoot, is a site for fostering an online community of researchers that apply data mining techniques in their scientific research. The portal provides researchers both access to and the ability to use over one hundred different ADaM mining algorithms exposed as services. The portal utilizes Talkoot modules to allow these researchers to perform tasks such as searching for distributed data, moving the distributed data to a computational resource, creating and executing data mining workflows, sharing workflows or full experiments with each other, and finally publishing science stories presenting their results backed by data and workflows.

References

Berendes, T., Ramachandran, R., Graves, S., Rushing, J., 2007. ADaM-IVICS: A software tool to mine satellite data. 87th AMS Annual Meeting, San Antonio, TX.

Graves, S., Ramachandran, R., Maskey, M., Keiser, K., Lynnes, C., Pham, L., 2007. Solutions for Mining Distributed Scientific Data. Fall American Geophysical Union Meeting, San Francisco, CA.

Ramachandran, R., S. Graves, T. Berendes, M. Maskey, C. Chidambaram, S. Christopher, P. Hogan, and T. Gaskins, 2009a: Visualize, Analyze and Mine Satellite Imagery Using a Single Tool 17th International Conference on Geoinformatics, Fairfax, VA.

Ramachandran, R., S. Graves, T. Berendes, M. Maskey, C. Chidambaram, S. Christopher, P. Hogan, and T. Gaskins, 2009b: GLIDER: A comprehensive software tool to visualize, analyze and mine satellite imagery. IEEE International Geoscience & Remote Sensing Symposium, Cape Town, South Africa.

Ramachandran, R., Graves, S., Rushing, J., Keiser, K., Maskey, M., Lin, H., Conover, H., 2008. ADaM Services: Scientific Data Mining in the Service Oriented Architecture Paradigm. In: Dubitsky, W. (Ed.), Data Mining Techniques in Grid Computing Environments. Wiley Publication.

Ramachandran, R., M. Maskey, A. Kulkarni, H. Conover, U.S. Nair, S. Movva, submitted 2011. Talkoot: Software Tool to Create Collaboratories for Earth Science. Submitted to *Earth Science Informatics.*

Ramachandran, R., S. Movva, and S. J. Graves, Ontology-based Semantic Search Tool for Atmospheric Science, *22nd International Conference on Interactive Information Processing Systems (IIPS), 86th American Meteorological Society Annual Meeting*, Atlanta, GA, 29 January-2 Feb, 2006.

Rushing, J., T. Berendes, H. Lin, C. Buntain, and S. Graves, "Spyglass: A System for Ontology Based Document Retrieval and Visualization", *The 22nd International FLAIRS Conference*, Sanibel Island, FL, 05/2009.

Rushing, J., Ramachandran, R., Nair, U., Graves, S., Welch, R., Lin, A., 2005. ADaM: A Data Mining Toolkit for Scientists and Engineers. Computers & Geosciences 31, 607-618.