A Workflow-Based Knowledge Management Architecture for Geodynamics Data

Jian Qin John D'Ignazio
School of Information Studies
Syracuse University
jqin, jadignaz@syr.edu

Suzanne L Baldwin
Department of Earth Sciences
College of Arts and Sciences
Syracuse University
sbaldwin@syr.edu

Abstract

A prototype project made up of Syracuse University earth sciences and information science researchers is exploring the cognitively demanding set of workflows required to make science advances in thermochronology and plate tectonics. The participants aim to understand the data practices and requirements that support publication in this research domain and use them to suggest knowledge management tool and standard development on two levels: 1) raw data file management and 2) knowledge management related to steps in sample and data processing. Enhancements in data management and curation functions are meant to standardize or automate processes in data verifiability, interworkability, analyzability, and interoperability, leaving researchers free to conduct higher levels of analysis and synthesis. This prototype would therefore offer critical lessons for EarthCube to increase the productivity and capability of geoscientists.

The Vision of EarthCube and the Missing Link

EarthCube is envisioned as "a knowledge management system and infrastructure that integrates all geosciences data in an open, transparent and inclusive manner" (NSF, 2011b). As a knowledge management (KM) system, EarthCube should accommodate science requirements regarding data curation for each stage of research, including data capture, processing, annotation, quality verification, cleaning, aggregation or extraction, transformation, analysis, and finally publication and preservation. This KM system would not be one-size-fits-all, but instead would have the architecture capable of supporting individual community members' needs. As an infrastructure, EarthCube will deliver many of the capabilities of a KM system in the form of services both in the sense of human-centered and interworkable data services. By human-centered, we mean the infrastructure can support effective knowledge creation through services that help minimize the time to make data analyzable. The interworkable data services refer to data tools and standards that help researchers work with different data types and formats for their science research.

This KM system and infrastructure vision will require geoscience data to have the following properties:

1. Verifiability: datasets should bear provenance metadata that allow researchers to trace them back to the raw data for quality control and data reuse purposes. The verifiability property ensures the validity of research and allows researchers other than the data owner to repeat the study using the same data.

2. Interworkability: datasets should contain sufficient metadata to facilitate data discovery, selection, aggregation or filtering, and reuse. The interworkability of data types and related metadata should be built to accommodate researchers generating data from the very beginning and throughout the research lifecycle.
3. Analyzability: datasets should be in a state that requires minimal data manipulation in order to proceed with science research. This property implies that the KM system prepares the data to be ready for analysis based on science requirements for one or more research communities. Such analysis-ready datasets would be of appropriate type and include necessary documentation or metadata delivered as part of infrastructure services.
4. Interoperability: datasets should conform to standards so that they can be communicated and processed by different systems and software tools. Such interoperability ensures that the verifiability, interworkability, and analyzability of data will not only transcend space and time, but also reach across the practices of the research communities that need to use or reuse them.

In addressing the data deluge challenges, much of the attention has been directed toward two broad areas. One is large-scale and high performance computing geared toward the science research aspect of data, that is, using software tools to analyze, model, simulate, and visualize data and create new knowledge (Hey, Tansley, & Tolle, 2009; NSF, 2011a). The other focus has been on the curation of science data, or the management of the end products of research in the form of the resulting data sets. While both high-performance computing and data curation are important part of the new eScience paradigm, there is a missing link in between these two areas: a focus on the services and tools that transform the data resulting from lab experiment, observation, or simulation into verifiable, interworkable, and seamlessly accessible forms that enable scientific discovery. It is not uncommon for scientists to be overwhelmed by the amount and variety of data their projects produce. They therefore spend more time getting data ready for analysis than proceeding with the research itself. Developing workflow-based science requirements regarding these data types and related metadata, and translating them into a KM infrastructure will establish links between large-scale, computing-driven discovery and data curation. This white paper proposes a workflow-based approach to acquiring and assessing data-related geoscience requirements in building EarthCube.

Cognitively-Demanding Workflows

Research on data-driven workflows is not new and has been reported in computing-intensive research fields, a.k.a., “big science” such as astrophysics (Brown et al., 2006; Deelman et al., 2010; Miles et al., 2008; Ludäscher et al., 2006). The “big science” workflows are computationally intensive and have little or no manual operation from data generation/collection to analysis. At each stage of this process almost all the processing and analysis tasks are specified and controlled by workflows that consist of a series of procedures, programs, and commands to be executed. In contrast to these “big science” implementations, most research involves tasks that require human cognition, manual operations, and interactions with computer and network technology to generate, collect, and acquire data before science analysis can begin. Examples of these equally data-intensive “workflows” include geoscientists’ work with rock, water, and soil sample collections. Data generated from this type of research often need to go through complicated processes to transform the samples into the forms suitable for lab experiment and analysis. Data collected from these lab tests and experiments are typically entered into spreadsheets, which accumulate into a large quantity of "small" digital files with varying data fields and annotations.

Using the thermochronology and tectonics research workflows will demonstrate the "missing link" problem as a contribution to the EarthCube vision of enhanced knowledge management. Thermochronology and tectonics research involves sample rock collection and a number of radiometric methods to determine the thermal and tectonic evolution of the Earth's lithosphere and planetary materials. The lab experiments generate data on minerals and rocks that reveal the timing and rates of plate boundary processes, the evolution of orogenic systems, and the timing of planetary processes. These data are usually stored in spreadsheet files or isolated databases and more often than not, include annotations. The data collected must be accompanied by careful documentation for quality control and track-back (to the rock sample) purposes. Unlike automatic data capturing devices that produce streams of bits and bytes, this process generates large amounts of "small" data files. To make data analyzable, the researcher spends a great deal of time in merging/ aggregating, filtering, or reducing data from hundreds or thousands of data files in order to produce a "master" dataset suitable for research analysis and modeling to show the phenomena of interest. Each datapoint has a complicated set of attributes and relations according to rock sample, processing lineage, and geolocation. The workflows generating these datapoints involve distributed effort from members of a research group who contribute a large amount of manual data capturing and manipulation as well as human cognitive capacity to perform. As a result, the science research is hindered by the cognitively-demanding and human-coordinated workflows.

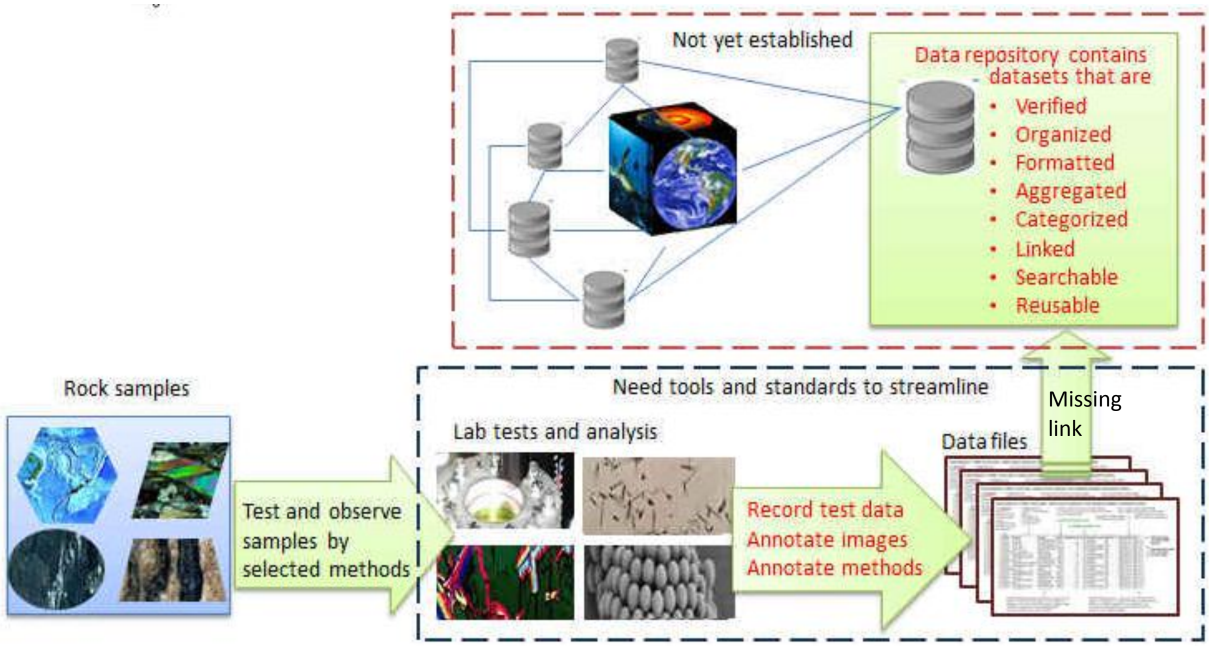

Figure 1. Thermochronology and tectonics research workflow illustration and

As illustrated in Figure 1, the lab tests and analysis apply multiple methods and instruments to obtain data on when and under what conditions (such as pressure and temperature) minerals and rocks form, and researchers record data from the instruments and make annotations as needed. While this portion of the workflow is supported by miscellaneous software tools in data capturing, a large portion of it is still done manually to be fail-proof and because there is a lack of software support for capture and annotation of the specific data. The cognitively-demanding workflows result in data files in various styles and formats, which create operational hurdles leading to expenditures of substantial amounts of time and effort to obtain analyzable datasets from the myriad data files. The state of these data files increases effort required for related research groups to conform to the larger geoscience data network and the EarthCube vision. The key to solve the data problems in cognitively-demanding workflows is to streamline manual operations by developing tools that can accommodate the science requirements of such workflows and establish automated links to the larger geoscience data network.

A Workflow-Based Knowledge Management Architecture

At Syracuse University, we are currently working on a knowledge management prototype for the thermochronology and tectonics research group's data management, using their NSF-funded Continental Dynamics Project as a prototype. This project aims to understand how the Earth's youngest ultrahigh pressure rocks have been exhumed to the surface in an active plate boundary zone in eastern Papua New Guinea. Their lab experiments determine when rocks and minerals form and link that data to structural (field) and tectonic settings to determine how the plate boundaries evolve in space and time. As mentioned earlier, the workflows in this research field can be characterized as cognitively demanding, and as such we started interviewing and observing the researchers to understand laboratory processes that capture data from samples of various types of rock derivatives (e.g., thin sections of rocks, mineral grains, sample locality maps, and structural cross-sections). We plan to develop step-by-step workflow and data flow details and relate them to information-organizing schemes used in the earth sciences. A thorough understanding of these workflows should prove invaluable for generalizing patterns and requirements for tool development.

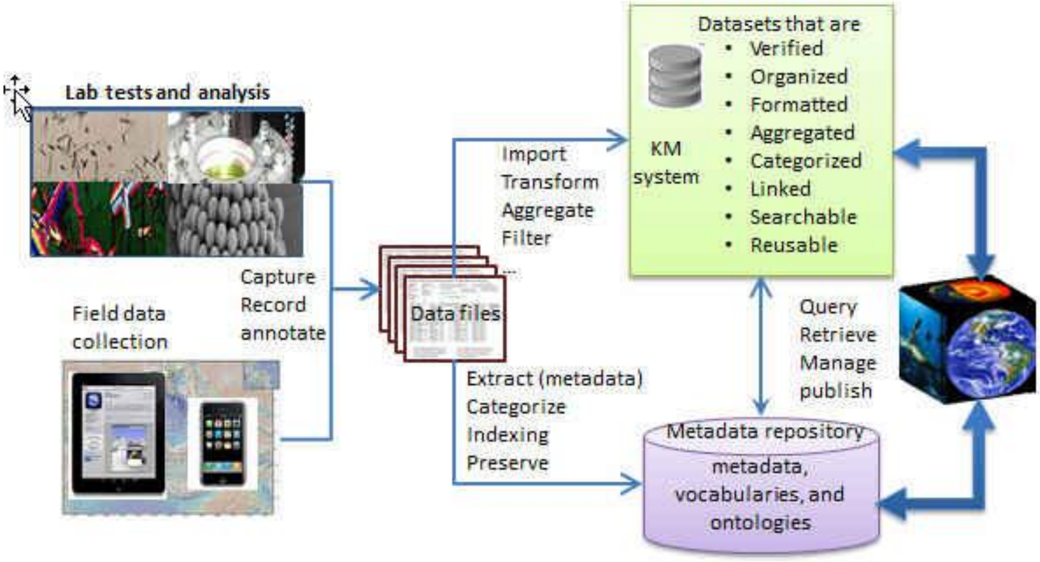

Figure 2. Illustration of workflow-based knowledge management architecture

A knowledge management architecture that supports the cognitively-demanding workflows of geoscientists according to the vision for EarthCube is shown in Figure 2. To meet the requirements for data verifiability, interworkability, analyzability, and interoperability, our prototype will build two major functionality areas: one for raw data file management and the other for knowledge management. The data management functionality focuses on streamlining the workflows for field and lab data collection and recording to enhance data collection effectiveness and verifiability that will allow researchers to trace back to the original data files and conduct quality control tasks. The knowledge management functionality encompasses a wide variety of data operations, including, but not limited to, the following:

- Enable data import, merge, and aggregation operations through a user-friendly interface so that researchers or graduate students do not have to deal with as many technical aspects;
- Enable provenance management for both the data and knowledge management functionalities;
- Develop metadata and vocabularies based on standards and science requirements;
- Establish linking and conformance mechanisms to support the data produced by individual research groups and uploaded to discipline-based repositories to be discoverable and reusable for interdisciplinary research;
- Develop a tablet-computer or smartphone application that supports digital data to be taken in the field, versus the use of paper forms; and
- Enable data sets to be imported into modeling programs such as Pecube, which uses finite element code to solve the transient heat transport equation in three dimensions, including heat conduction, advection and production as well as the effect of a finite amplitude, time-varying surface topography (Braun, 2003).

In summary, the goal of this proposed prototype is to develop tools and standards that can support geoscientist practices in field data collection and lab test data recording that leads to more powerful research discovery. This development is based on an in-depth study of the workflows of an earth sciences research group and their research field of thermochronology and plate tectonics. Initial findings of our prototype-development project show the cognitively-demanding nature of this research field. Continued exploration of the research group’s science workflow requirements in order to develop or apply appropriate tools and standards to data curation functions should serve to support these scientists by allowing them to concentrate on higher level research analysis and synthesis tasks. Such a prototype project offers critical lessons for the broader EarthCube initiative.

References

Ludäscher, B., Altintas, I., Berkley, C., Higgins, D., Jaeger, E., Jones, E., Lee, E.A., Tao, J., & Zhao, Y. (2006). Scientific workflow management and the Kepler system. *Currency and Computation: Practice and Experience*, 18(10): 1039-1065.

Deelman, E., Berriman, B., Chevenak, A., Corcho, O., Groth, P., Moreau, L. (2010). Chapter 12: Metadata and provenance management. In: A. Shoshani and D. Rotem (Eds.), *Scientific Data Management: Challenges, Existing Technology, and Deployment*. CRC Press. Available at: http://arxiv.org/ftp/arxiv/papers/1005/1005.2643.pdf

Braun, J. (2003). Pecube: A new finite-element code to solve the 3D heat transport equation including the effects of a time-varying, finite amplitude surface topography. *Computers & Geosciences*, 29(6): 787-794.

Brown, D.A, Brady, P.R., Dietz, A., Cao, J., Johnson, B., & McNabb, J. (2006). A case study on the use of workflow technologies for scientific analysis: Gravitational wave data analysis, in I.J. Taylor, E. Deelman, D. Gannon, and M.S. Shields (Eds.), *Workflows for e-Science*, chapter 5, pp. 41–61. Berlin: Springer-Verlag.

Hey, T., Tansley, S., & Tolle, K. (Eds.). (2009). The Fourth Paradigm; Data-Intensive Scientific Discovery. Redmond, Washington: Microsoft Research. Available at: http://research.microsoft.com/enus/collaboration/fourthparadigm/

Miles, S., Groth, P., Deelman, E., Vahi, K., Mehta, G., & Moreau, L. (2008). Provenance: The bridge between experiments and data. *Computing in Science and Engineering*, May/June: 38-46. Available at: http://eprints.ecs.soton.ac.uk/20874/1/cise2008.pdf

NSF. (2011a). Advisory Committee for Cyberinfrastructure Task Force on Grand Challenges Final Report, March 2011. Available at: http://www.nsf.gov/od/oci/taskforces/TaskForceReport_GrandChallenges.pdf

NSF. (2011b). EarthCube guidance for the community. Available at: http://www.nsf.gov/pubs/2011/nsf11085/nsf11085.pdf