A Novel Research Approach to Enhance Research Group-level Science Data Management

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

ABSTRACT
This poster will report an alternative approach compared with large-scale collection, technology development, or discipline informatics projects by exploring the enhancement of science data management at the research group level. An academic-year-long project involving both information and technical analysts demonstrated that flexible introduction of IT in the form of a relational database solution presents an opportunity to incorporate data-related practices, behavior, and schemes into a tool. The project team used cognitive work analysis and participatory design research methods to sensitively analyze research practice and then design and introduce a prototype tool. Iterative analysis and design cycles captured and incorporated the complex practices of researchers who struggle with the accumulation of samples and related digital data files from years of research. Design research processes will be analyzed according to project developments and research group reaction to the proposed data management tool design and preliminary attempt at implementation.

Keywords
Science data management, cognitive work analysis, participatory design research, e-science, relational database scheme

INTRODUCTION
As part of the IMLS-funded eScience Librarianship (eSLib) program at the School of Information Studies (iSchool) at Syracuse University (SU), researchers and students conducted a participatory design effort with an SU research group of earth scientists in order to address their data management issues. The project, informally known as “GeoDataMan,” ran the duration of the 2011-2012 academic year and fulfilled an educational objective of eSLib to give two students, enrolled as fellows in the iSchool’s MSLIS program, real-world experience in helping scientists manage their data. A master’s student in the iSchool’s Information Management program skilled in database design and programming also participated as a team member.

The GeoDataMan project, run by the eSLib’s graduate assistant with oversight and involvement by the eSLib principal investigator, also accomplished a research goal to understand how participatory design processes contribute to science data management at the research-group level. These processes included interviews, meetings and lab tours with the geology research group members who study the conditions (e.g., depth, temperature) in which minerals and rocks form, and their age, a research subfield known as thermochronology and tectonics. It also included an information analysis of the scientists’ environment, articles, websites, and datasets that led to the creation of multiple GeoDataMan artifacts such as tables and scheme diagrams, culminating in a relational database prototype designed to capture, store, and allow queries on the research group’s data.

APPROACH AND METHOD
Science data management (SDM) has received sufficient attention to warrant reviews of the state of the art in the curation practices to capture and preserve the digital forms in which these data exist (Ball, 2010). The approach to SDM that appears in journal literature or technical reports may be viewed from three distinct vantage points: either (1) an effort to assemble a large collection of data sets that serve a particular domain of science research (Baker & Yarmey, 2009); (2) a technical development program designed to facilitate a whole set of capability involving science data (Howe & Cole, 2010); and (3) the application of information systems to enable investigations by a large number of researchers working in a particular discipline such as earth science, developed with a “geoinformatics” project label (Sinha, et al., 2006).

An alternate approach is to interrogate the ability of information technology (IT) to represent and host digital data that is created by scientists and their research assistants. The GeoDataMan project was an attempt to follow this approach, with a goal to understanding a research group’s data practices, and apply such understanding to the design of an information system tailored to a research group’s practices related to data creation, reuse, and interoperability (Cammarata, et al., 1994). The IT applicable for such a project is a relational database, either open-source or proprietary, that can be
relatively quickly programmed to store either data or metadata and features a web-hosted interface, complete with access controls.

The GeoDataMan group formed a campus partnership with a research group of earth scientists, incorporating experience in eScience Librarianship that close association is necessary to understand the complex relation between diverse practices, instrumentation, computers, and data types that are input or output during research work (Wallis, et al., 2010). The group incorporated qualitative methods of interviewing and lab visits to elicit and analyze scientists’ behavior related to their data practices (Cragin, et al., 2010). The GeoDataMan project contributed to SDM-related research because of the effort to use this understanding of data-related practice and apply suitable database and data curation skills that result in an SDM tool designed particularly for the “client” research group.

An SDM-focused tool has two aspects that significantly relate to scientists’ practices at the research-group level. The first is that it is composed of a strong information framework related to the work environment in the form of a database scheme of entities and their attributes, and the relations possible between such entities. The GeoDataMan group employed Cognitive Work Analysis to address this aspect and define the entities and attributes of thermochronology and tectonics data, as well as their semantic or practical relations (Albrechtsen & Pejtersen, 2003). The second is that from the multitude of data-related entities associated with the research group, a fit must be established between IT capability and the right set of data elements defined and stored that would help the scientists’ in their day-to-day practice, and thus improve the chance of the IT-based tool being of real use. Similar information system design problems use a participatory design research strategy, so the GeoDataMan project adopted such processes to address this aspect (March & Smith, 1995).

The GeoDataMan project team met with the earth scientists several times throughout the academic year in meetings lasting up to two hours; conversations were digitally recorded. These recordings proved helpful in reviewing the abstract details, processes, and vocabulary of two senior scientists and a postdoctoral researcher discussing their research and associated operations across the full range of their rock samples and data files. In addition, artifacts such as photographs from a lab tour and sample data files were collected and shared among the GeoDataMan team.

Because of the richness of the “client” information, and the desire not to tax our scientist partners in having to take GeoDataMan staff through basic education on essential but elemental aspects of their research, we pursued a divide-and-conquer strategy. Two eSLib fellows performed an “information analysis,” where each took on a separate research method used in thermo-chronology and tectonics. The fellows examined the method’s distinct set of instruments and practice steps related to data production, integration, and analysis that comprised a class of workflows in the research domain. A third team member decomposed an existing web-based data repository that stored data from articles in this research area to formulate an initial database prototype. The entire team repeatedly came together to integrate their separate analyses and iterate the design of a database scheme. As design research methodology dictates, the GeoDataMan team displayed their generated scheme design and a more narrowly-focused

![Figure 1: Generalized thermochronology and plate tectonics workflow](image)
database to their client research group and received feedback on the accuracy and fit of these possible solutions.

**PRELIMINARY FINDINGS**

The GeoDataMan investigation into the SDM issues of a earth science research group provided a stark example of the negative impact of a research group’s sample collection and lab experiments accumulating relatively small in bit-size but numerous digital data files. The GeoDataMan team witnessed examples of how this local “data deluge” hampered progress because redundant or taxing manual work was necessary. For example, researchers had difficulties in finding the data that matched exact criteria related to momentary but recurring interests. The research group members had limited control of versions of data files, and limited ability to merge or segment data by inherent characteristics. While data quality control was critical for the earth scientists’ analyses, the lack of an integrated data management tool made quality control cumbersome. This multitude of data files and temporary “data views” in the form of proliferating sets of spreadsheets, however, had a consistency in their creation and manipulation governed by tool, method, or analysis phase used as the researchers proceeded to journal article publication.

At each stage of a generalized workflow based on the GeoDataMan information analyses (Figure 1), the scientist’s tasks required cognition, manual operations and interactions with samples, instruments, and computers. These steps generated data that provided the scientists’ an opportunity to compose findings based on analysis and integration across multiple, individual workflows. It at first seemed an overwhelming challenge to the eSLib fellows to analyze these workflows and understand the data entities and attributes, and their relations, due to the abstract definitions and complex procedures of this research domain. The limited access to the scientists for ready explanation also proved daunting.

Effort to analyze the data-related representations and build an SDM tool helped the GeoDataMan group understand where data transformations followed specific workflow paths, whether routinely merging or split into isolated practices marked by unique definitions. eSLib fellows learned to depend on their analysis of explanations and representations of research methods that governed the research group’s workflows. They based their information analysis on the research domain’s years of method development or adoption that appeared in earth science journal articles, monographs and textbooks, research reports and in domain-specific websites—information resources created as part of the science communication process.

An IT focus for this information analysis provided scope for the GeoDataMan team’s effort to extract enough information to frame the SDM tool but prevent their getting lost or attempting to become experts in these methods. Design research processes helped them artfully accommodate a set of the science practices they saw into a modular data management strategy represented in a relational database design. They also showed leadership and initiative by reflecting back to the earth scientists’ elements of their research work in a rich conversation about the ability of an SDM solution to fit the client research group’s work.

**CONCLUSION**

The main result of this design research project was the design of a data model to enhance SDM based on a research group’s workflows and data flows. It focuses on collection—query behavior afforded by the SDM tool is an additional project phase because of the associated cognitive, task, and science practice dependencies. The data model and implementation resulting from the GeoDataMan project is not only a technical architecture for a research group’s data management solution but also establishes an alternative approach to solving data management problems in “small science” fields. Further analysis is planned to understand the deployment of cognitive work analysis and participatory design research processes matched to project progress. This project’s findings apply to a research group working in a particular domain of earth science, but are likely applicable to research groups working in other fields such as biomedicine, ecology, and environmental science.

**ACKNOWLEDGMENTS**

The eSLib project PI and staff are grateful for the initiative and contribution of Christopher Turner, Trisha L. Adamus, and Bin Zhu, as well as the funding provided by the Institute of Museum and Library Services (IMLS) #: RE-05-09-0061-09.

**REFERENCES**


Howe, B. & Cole, G. (2010). SQL is Dead; Long Live SQL: Lightweight Query Services for Ad Hoc Research

