BUILDING A COMMUNITY-BASED DATA NETWORK FOR GEOTHERMAL ENERGY

Walter S. Snyder, Geothermal Data Exchange, Department of Geosciences, Boise State University, Boise, Idaho; Joseph N. Moore, Energy & Geoscience Institute, University of Utah, Salt Lake City, Utah; David D. Blackwell, SMU Geothermal Laboratory, Department of Geological Sciences, Southern Methodist University, Dallas, Texas; Tonya Boyd, Geo-Heat Center, Oregon Institute of Technology, Klamath Falls, Oregon; Roland N. Horne, Stanford Geothermal Program, Department of Energy Resources Engineering, Stanford University, Stanford, CA; Lisa Shevenell, Nevada Bureau of Mines and Geology, University of Nevada, Reno, Reno, NV

ABSTRACT

There are many issues associated with the development and sustainability of a network of data sites and databases hosted by academic-based groups. Some of these are technical, most are nontechnical issues. These academic institutions have always had the dual missions of conducting research on geothermal systems while educating the next generation of geothermal professionals and researchers. Now, a third role is emerging, that of data stewardship as it applies not only to research and education, but also: 1) as a tool for industry as they push forward with delineating and producing geothermal resources, 2) for state and federal agencies to help them meet their missions and mandates, and 3) as a tool to inform the public on the importance of geothermal energy.

The basic notion of a data network is that several data sites come together to collaborate on acquiring particular suites of data and making them available to the larger user community. Over the last ten years, there has been a growing awareness of the importance of better data management; indeed both Congress and the White House continue to strengthen the bipartisan goal of free and open access to all data created by the federal dollar. For academic-based data sites, there are many challenges to building and sustaining an effective data network. The first challenge for a network is to agree on the system standards for sharing and providing access to the data among the data sites, the nodes, on the network. Several international and national groups have developed global standards that can be utilized, but the key issue is on the specific implementation of these as system standards for the network. The bigger challenges are operational and reflect social-cultural and political realities. The conclusion is that first and foremost the focus must be on the geothermal user community. Then a successful network must operate under the principles of openness, collaboration, flexibility and a willingness to change. The latter is critical as the developers and the community being served become more knowledgeable and involved, as technologies evolve, and as opportunities for sustainability come and go. As long as the academic groups and the interested federal and state agencies are willing to collaborate there should be few barriers to creating the envisioned dynamic system.

INTRODUCTION

Significant growth in contribution of geothermal to the international energy portfolio requires reducing the risks and cost of defining resources, characterizing new classes of larger energy resources, optimizing management and expansion of exploited geothermal fields, expanding direct use of geothermal, and ensuring a path for technology growth into the future, in particular providing the science and engineering basis for conventional and enhanced geothermal systems (EGS). All of this is predicated on an enhanced knowledge, and knowledge requires accessible data. This summary statement comes from an unpublished proposal (W.S. Snyder and J. Moore, co-PIs) and is summarized on the National Geothermal Data System website (NGDS, in which all co-authors participate; www.geothermaldata.org). In this paper, we do not address the NGDS specifically, although most of what we say is applicable. Rather, we are addressing the fundamental challenges for academically based groups involved in creating, maintaining, sustaining, and expanding such a network, utilizing the data needs of geothermal energy as an example. We certainly hope that the insights we present here will be incorporated into the NGDS as it moves forward. We are also fully aware of the importance of working with federal and state agencies on these endeavors and of international collaborations - we are doing both. But here, we focus on the issues associated with bringing together this group of core academic institutions. A network can be envisioned as an Internet-connected series of nodes (data sites), that allow for a common approach to finding data among the linked sites. Each of the co-authors' groups have for years collected and provided data to researchers, industry, state and federal agencies, and the public and this collaborative approach extends the reach and effectiveness individually and collectively. The issues discussed here reflect our collective experiences.

DATA STEWARDSHIP - WHY YOU CARE

Over the last ten years there has been a dramatic increase in awareness of the need to fully manage data generated by research and development activities, industry and federal and state agencies. This long education process perhaps has not yet peaked, but the realization that data are the underpinnings of science and engineering, the basis for investment decisions, and that they are crucial for land and natural resource management has been noted and documented by the National Science Board (NSB, 2005), the National Academy of Sciences (NRC, 2002, 2009), and emphasized by Congress and the White House (e.g., Interagency Working Group on Digital Data, 2009; OSTP, 2009; and http://www.ostp.gov/cs/issues). This awareness is continuing to grow with the advent of federal agency data management plans and requirements, the awareness from major publications of the importance of data, continued discussion by the Federal Interagency Working Group on Digital Data and in a number of National Research Council reports, as well as in the general literature (e.g., NRC, 2002; Atkins et al., 2003; Atkins et al., 2011; Hey et al., 2009; Nature Editorial, 2009; NRC, 2009; and many more). We seem to be moving at a faster pace. This is a good thing. Both Congress and the White House continue to strengthen the bipartisan goal of free and open access to all data created by the federal dollar. However, we as an industry, science, and nation have not done an adequate job of capturing and providing these scientific data to users (researchers, industry, state and federal agencies, and the public) and not just data produced by federal funding.

Many of the data critical for this expansion of geothermal energy are inaccessible - they are beyond the reach of those who could use them. A Department of Energy DOE report by Deloitte (2008, pg. 27) concluded that: "A study conducted in 2000 for NREL (Entingh, D., 2002) revealed that over a 25-year period, numerous geothermal research efforts were conducted with state and federal funding and that the analysis and information contained in those research documents are difficult to access. That same study cited that much geothermal resource attribute data also exists but is distributed among numerous locations and often stored in boxes, without any data index or organization. Even these identified data represent a small part of the overall data that exist, but is inaccessible and that would significantly help the efforts to expand geothermal's portion of the nation's energy portfolio if we could find and access them."

But the issue of data stewardship goes beyond geothermal and DOE - it is a general problem that cross-cuts many disciplines and institutions. Each person, be they a researcher, employee of a company or a federal or state agency, needs to become more aware of the long-term value of the data they generate through their activities - to become better data stewards. For researchers, no longer is it sufficient for them to document their work by only publishing a paper, even if a supplemental data table is included. For companies, data management is an issue of retaining knowledge, the corporate memory, making better business decisions, and being able to do a better job of attracting outside investments. For agencies the impetus for better data stewardship can be a mixture of what drives both researchers and industry, but also the fact that they are the public's stewards of data generated by their tax dollars. It is not sufficient to think of data management only in terms of datasets and their associated papers. We need systems where all data associated with all research can be accessed in their most granular, discrete form while maintaining the attribution of each bit of data to its original author. These data must be openly accessible, once they are public, but held privately during a publication moratorium period. We need to have seamless links from the databases to publications. This will allow future users to easily move from the published paper to the data and metadata behind the publication and just as easily utilize these data in their ongoing research as well as give researchers citation credits for their efforts at data stewardship.

THE VISION

For geothermal energy, a data network as a system needs to capture the full geologic, geophysical, and engineering context of geothermal systems on scales ranging from regional to the individual well bore to the thin section and microscopic scales. Thus the system must be able to handle physical, geophysical, geochemical and a host of other data for use by scientists, engineers, project managers, investors, researchers, and others. In addition to supporting the science and engineering aspects of geothermal resources and associated research, the system would provide the basis for financial investment risk analysis. It will also support state and federal agencies with land and resource management missions and serve as an interface to the public and decision-makers. Finally, it can and should be designed to contribute to enhancing the education pipeline and diversity for people entering the geothermal industry. In summary, it is far better to under-populate an expansive data system than it is to rebuild a narrowly designed one. Hence, the ultimate system must meet the breadth and depth of needs as we can see them now and that is designed to efficiently and effectively expand and migrate into the future as the needs, visions, and technologies change. It cannot be built all at once, but having a clear roadmap of where we want to be is critical to the network's long-term success and viability.

SOME BASICS

Data Types

There are many ways to describe the types of data that must be accommodated, but the baseline distinction is as follows:

- **Data resource:** a generic term for all digital files that can be stored at a data site.
- **Data product:** includes preformatted text documents, photos, diagrams, datasets, videos and viewable maps. Metadata may or may not be included or may be incomplete.
- **Datasets:** a type of data product where discrete data are provided, typically in spreadsheets, sometimes word processing tables. They are "products" in the sense that they are usually pre-populated and preformatted with data selected by the author, not the user. Metadata may or may not be included or may be incomplete.
- **Discrete data** are the "base" or "raw" data that populate the tables and fields of a database; these include data and the metadata that describe the data.

In addition, all data products and most datasets are "static"; that is they reflect the content views and filters of the authors who created them and cannot (and should not) be modified. Conversely, some datasets and all discrete data are "dynamic" in that the content of any grouping of data may change with time as more data are captured into the data system. The concept of a “dynamic dataset” is important. A dynamic dataset (which can include or solely be comprised of geospatial data) is one whose structure is defined by an author or user, data are pulled from a structured database (or data warehouse), and are updated periodically from the database, hence “dynamic” (for example, time series data from a remote sensor).

Metadata

Metadata are “...‘data about data’, or more explicitly structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage a … resource” (NISO 2004), be that a digital document such as a report in Word format, a spreadsheet of chemical analyses on a rock sample, a photo, a map, etc. The definition or at least the application of metadata can become blurry because what metadata is to one person may be data to another. The easiest way to think about this is to not overly worry about the distinction between data and metadata, and ask and answer the question: “Do I have sufficient data to totally describe the feature I’m dealing with?” - for example, a chemical analysis of a rock. In addition, metadata are a means of allowing others to find data - as long as enough metadata information is provided. What are the results reported as (elemental, oxide, etc.)? What are the measurement units? What are the errors and what type of instrument was utilized? What standards were used? Etc. These metadata are critical to document the quality of the data - without them, allows the assumption that the data are of lower quality. Each metadata element has to have a definition that is more formal than those just listed so others (or machines) can understand what is meant. Therefore, metadata are important. Also each piece of discrete data has to have a definition associated with it - something that describes that data point, be that in a cell in a spreadsheet or a field in a structured database. Finally, the data are not random, but have relationships among them; for example the SiO2 as a type of chemical analysis (analyte) has to be associated with a sample (think sample number) as well as with its value (e.g., 53.2) and that value with a unit (e.g., %).

Extensible Markup Language (XML)

XML (eXtensible Markup Language) has been an indispensable part of moving data around on the web including the catalog and thematic web services mentioned below. It is a set of rules and guidelines for describing structured (or semi-structured) data in plain text, standardized by the World Wide Web Consortium (W3C). It is used to create a text-based file with “tags” that describe and provide structure to the data, which together with the schema (the ordered relationships among the data) make the document machine readable, understandable and parsable so data can be extracted by the user’s computer in an application, such as a browser. It requires that each data element have a specific name and definition. The power is that if an XML schema is adopted as a standard, say for well logs or bottom hole temperatures, then it allows these relevant data and metadata to be mutually shared across nodes on a network and with any other global site provided they accept that particular schema. Also, such standardized content promotes easy data mash-ups by the user who may find data at various sites and desire to compile them into a single data set. Finally, it should be noted that newer mechanisms for data exchange on the Internet continue to evolve, e.g., JASON, REST, etc., but the toolsets and standards for these have not yet matured to the same level as XML and XML schemas.

User Focus

The major lesson learned to date is that technology alone cannot drive the creation of a network - rather, more attention must be paid to the users, those that generate and use the data and have little or no interest in the technologies behind the data systems. IT mechanisms should influence, but not dictate how data are acquired, what data are acquired, or how those data are tagged for archiving. For example, web services, linked data mash-ups, and Resource Description Framework (RDF) documents are wonderful tools that have sprung from the notion of the semantic web and efforts of the World Wide Web Consortium (W3C) and Open Geospatial Consortium (OGC) to make data machine readable and easier to find over the internet. These are very useful tools for data discovery and access, but when they are used to dictate to users and/or data sites what data and metadata they need to capture and how they should work with and present those data, without regard to the goals, needs, resources or time frames of those users or data sites, then the process has been reversed and IT is dictating technical expectations rather than responding to scientific and engineering community needs. However, the thesis here is that all technologies, including those of the semantic web, must be considered, developed and implemented within the context of their impact on real world social-cultural-political-economic frameworks. In effect, a mega-use case scenario. These use cases should not be constructed from the view or vision of a perfect world, but with respect to their impact and acceptance by real people, organizations and institutions. One advantage of an academically based network is that it is rooted in the user community and thus inherently does a better job of understanding the views, workflows, and needs of the people who comprise the geothermal community.

THE NETWORK

For academically based data sites, the challenges for building and operating a collaborative network are many, but not insurmountable. For example, seamless linkages of data to analysis and visualization tools need to be provided, in particular high-level modeling programs and required computational resources. When dealing with research results and data involving industry partnerships, moratorium and proprietary data must be handled carefully and securely. At the same time, it must be made easy for users to discover, aggregate and synthesize data in ways that allow them to focus on the analysis of these data rather than on finding and compiling them. Additionally, research-level data must be better utilized within the education enterprise to train and attract our next generation of geoscientists and geoengineers.

Network Operation

Inherent in the definition of a data network is that it provides more than just links pointing to other websites, the type of URL link you find on most websites that direct the user to other sites of possible interest. The underlying goal of any network is to interoperate at some level that makes the finding and sharing of data easier by the nodes on the network and/or outside users. This can be thought of as two levels of data service in networks: 1) "data sharing"; sharing data among the nodes on the networks, and 2) "data access"; providing the outsider user with single point access to data from all nodes at once. The Environmental Information Exchange Network (www.exchangenetwork.net), which has been operating for over seven years, is an example of the latter, and focuses on the needs of each node on the network, and then each node serves its own customer base. The developing NGDS is an example of the second type where each node remains the steward of the data it hosts, and a common catalog of data that each site hosts is made available for users to search and discover the data of interest. The data access type of network provides access to data through two basic methods: the catalog and digital library, and the thematic web services.

Digital Library:

A digital library includes a central catalog which includes the metadata index of its data resources and mechanisms for user retrieval of those data resources. These metadata include a specific "Uniform Resource Identifier" (URI) that provides the unique internet address of the specific data resource so that users can find and download the resource. Thus, the catalog facilitates user discovery and access to the specific products of interest. One need expressed by the geothermal user community is to provide the ability to be able to access a broad spectrum of data resources, preferably through one search location. The catalog can do just this by providing the user with the ability to search the central catalog for resources held in data repositories at various nodes throughout the system. The user should be able to search by text string and for those data resources with geospatial metadata, through a map browser. Users should be able to download all discovered items by standard methods.

Thematic Web Services:

Web services provide mechanisms to move data over the Web through a proscribed set of technologies that are an outgrowth of a World Wide Web Consortium (W3C) initiative; they are now, for the most part, commodity items. The nomenclature "thematic web services" is used to distinguish the process of using web services as user accessible, pre-defined search and data retrieval mechanisms from those IT operations where web services per se are used for a number of background operations. The two should not be confused. Thematic web services provide users with pre-defined data products and datasets and contrast with ad hoc search and data resource retrieval. We have developed several web services as part of our ongoing work on the NGDS.

The key to successful implementation of such web services is to work with the user community to identify those services that various segments of the community would find useful. A "web services listing" should be part of the system catalog to provide a central point for users to survey the available web services and select which ones they might want to utilize. Some thematic web services will be specific to a particular node, and the URL will reflect that. If the web service pulls data resources from multiple nodes, this should be transparent to the user.

STANDARDS, PROTOCOLS AND BEST PRACTICES - A MOVING TARGET

It is incumbent on any network, indeed any data site, to compare and assess the relevant standards, protocols and specifications being worked on and/or implemented by a variety of national and international groups. For the geosciences, these groups include the Marine Metadata Interoperability Project (MMI; www.marinemetadata.org), EarthChem (www.earthchem.org), the U.S. Geoscience Information Network (USGIN; usgin.org), CUAHSI Hydrologic Information System (HIS; http://his.cuahsi.org), Canadian Well Logging Society (www.cwls.org), Energistics (www.energistics.org), CGI (Commission for the Management and Application of Geoscience Information of the International Union of Geosciences), Dublin Core (an implementation of ISO standards (see below) for data product metadata, www.dublincore.org), the NGDS (www.geothermaldata.org), and others. These groups in turn are assessing and adopting various standards, protocols and specifications sanctioned by organizations such as the Open Geospatial Consortium (OGC), World Wide Web Consortium's (W3C) web service standards and specifications, International Organization for Standardization (ISO; in particular ISO 19115, ISO 19139), Federal Geographic Data Committee (FGDC), the North American Geological Map Data Model (NADM), and others.

Operational Issues

In an ideal world, the adoption of these standard specifications and protocols by a data network would be easy. However, in the real world it is not for several reasons. First, and perhaps most confusing are conflicting standards and/or differing implementations of the standards from two or more standards groups. This quote from Wikipedia on standards for library documents exemplifies the problem:

> "Standardization for library operation has been a key topic in international standardization (ISO) for decades.

Standards for metadata in digital libraries include Dublin Core, METS, MODS, DDI, ISO standard Digital Object Identifier (DOI), ISO standard Uniform Resource Name (URN), PREMIS schema, Ecological Metadata Language, and OAI-PMH." (http://en.wikipedia.org/wiki/Metadata).

Second, most of these standards and/or their implementations reflect IT development, not necessarily what the scientific user community needs or the way it works and thinks. What group is in the best position to assess this latter question - one outside of the community or one from within the community? And third, some of these standards, while helpful technically, may compromise the content and therefore the utility of data. Again, what group is in the best position to assess this particular question? Specific issues relevant to the discussion on the development and sustainability of a geothermal data network are content models and their resulting XML schemas.

Content models

Content models capture the data and metadata content needed to describe a specific geological, geophysical, engineering, or other feature or entity, for example a well log. Think of these as the column headers in a spreadsheet where the follow-on rows denote particular instances of a feature or entity. Content models are important in part, because they can be used directly to develop thematic web services and/or the data from them can be extracted and aggregated into a database that serves as the basin for web services and other search and download operations. The problem is that the content models can vary widely depending on the community they are meant to serve. Are petroleum well log standards now being promoted by Energistics (www.energistics.org) the ones needed or used by the geothermal community? How do you handle legacy well logs that do not fit those evolving "standards"? Choosing content models can be problematic because they raise several questions. Whose definition, whose model do you adopt? Does it reflect the community of users it is meant to serve or does it have some other purpose? Is it so complex that it will not be used? Is it too simple that it does not provide sufficient description? Who makes these decisions?

Content models to XML, Catalogs and Web Services

The user may well see content models in the form of spreadsheets with pre-defined fields they fill out ("loaders/templates") and give to a database, but they won't see the XML code that is extracted from them and that is used in data storage, discovery and sharing; the XML that is the heart of the data network's catalog and web services (see above). The power is that if a particular XML schema is adopted as a standard, it allows data to be read and translated by any system and therefore the data are more easily shared, compiled and understood. The challenge lies in deciding whose XML schema is adopted for a particular subject.

XML schemas of particular interest for geoscience and geothermal include: GeoSciML (a mark-up language developed initially for geologic maps, but being extended for mineral deposits and mining and other geologic entities; www.geosciml.org), International Geo Sample Number (IGSN; formerly SESAR; global sample number standardization; www.igsn.org), CUASHI's Hydrologic Information System (HIS) WaterML (focuses on surface water hydrology, but will be extended for subsurface hydrology in the future; http://river.sdsc.edu/Wiki/ WaterML.ashx), EarthChem XML (targets geochemical data, http://www.earthchem.org/developers), and USGIN (content models only), and others.

The fundamental problem with the content model-XML couplet is that every content model-schema requires singular definitions for each data element and a set schema and associated ontology of how the data are interrelated (that is the "knowledge" of the data structure). A much longer discussion on knowledge and ontologies is needed to fully explore this problem, but in short, the problem revolves around the fact that data and knowledge are not the same thing. Linking of data and making it easier to find data and the pre-wired relationships among data elements may contribute to knowledge, but 'knowledge' is much more than that. Data, while comprising the foundation of knowledge, are an insufficient measure of it. Furthermore, and perhaps most important for the practical pursuit of bringing more geothermal energy online, if everyone is forced to use one definition for each word and to link those words in a single type of sentence structure, the construction of new knowledge is actually curtailed. While making it easier to find data online, a rigid content model-XML couplet can make it more difficult to innovate with those data in the real world of geothermal energy which is complex and incompletely understood.

The issue of content models and their XML schemas is complex and important for data systems, in particular geothermal data, because so little of the needed data have yet been captured by any data system and we will have to rely on the community to help populate the databases; indeed the users may be required to do so via new federal funding policies. For the foreseeable future, the answer to the question of whose XML schema to use lies in the operational "data sharing" approach, championed by the Environmental Information Exchange Network mentioned earlier. This approach utilizes translation and data interchange templates that allow nodes to share data. It also provides for the construction of a common catalog for finding data from the network as a whole. This approach allows data from an entire network to be shared with that from another network without forcing each node or network to operate in exactly the same way. Over time, and with international collaboration, there will likely be convergence toward more shared implementations of specifications and protocols, but that cannot be a forced operation. Uhlir, P.F., et al. (2009) and other studies have captured this natural flow and emphasized that convergence will happen naturally, over time, if it is allowed to happen. The point here is that this convergence must be allowed to happen naturally for all data systems and in particular for geothermal data. If barriers are created to this convergence process, then a true community of best practices and data sharing will not develop and cannot be sustained.

Summary

Experience to date provides three lessons. First, a practical operational question is whether or not the user will fill out all the fields of a complex content model. If they won't, then what? Structuring the content models into minimal/required, recommended, and optimal data helps. The optimal (and therefore most complex) level is the most desirable, but the minimal level will at least allow key legacy data - and also newly generated data - to be captured by a data system and therefore not be lost. Second, translation templates need to be used for sharing data with differing XML definitions and schemas both within the network and among networks. Finally, a true, open and global dialog needs to be developed on the issues of standards, protocols and best practices among groups interested in the problem for the geosciences, in particular for geothermal. This process must recognize that different groups have different missions and mandates that must be accommodated.

In summary, it is imperative for the network to assess, adopt and implement standard specifications and protocols in a deliberate and considered manner. That takes time, and thus a philosophy of being flexible, and an architecture that allows change. This flexibility includes the way the specifications and protocols are implemented. This is particularly true when tying together existing data sites into what become nodes on the network. Metadata and data content models and their resulting XML schemas and vocabularies are key examples of where the implementation of "standards" can be problematic, but this does not present insurmountable problems if short-term solutions are not forced upon a situation that is inherently a long-term process.

NONTECHNICAL ISSUES

Building a viable, collaborative data network revolves around nontechnical issues more than it does technical ones. As noted, for an operational data network, the system members technically interact through a common set of standards, protocols and specifications for information discovery and interchange. The system also provides a framework to manage, coordinate and maintain system activities and products. The system may provide a publically recognized, central place to begin the search for geothermal data (e.g., for the NGDS, the www.geothermaldata.org site), however, the intent of a distributed system is to allow users to begin their search at any point in the system. This latter point is important because it maintains the equal standing of all nodes on the network.

There are many fundamental issues that must be addressed before we can achieve the vision of a seamless, integrated data network, including:

1. Developing a process to reach agreement on the standards, protocols, technical specifications, etc. required to share data between systems.
2. Minimizing the changes that the established system data sites need to make to participate
3. Recognizing and accommodating the fact that a single solution may not be achievable (financially, technologically, culturally, politically, etc.) so this cannot be a basis of system functionality.
4. Understanding that systems need to be responsive to their users, so agreements about technical issues need to be adaptable to the needs of the users.
5. Providing flexible, clear and concise operational procedures as well as specifications that make it possible for other data sites to join the network in the future.

Finally, an underlying design criteria for a distributed system is that all associated data sites are and will remain independent entities with their own missions and mandates, and will receive full credit for the data they serve regardless of the point of access for these data.

SUSTAINING THE NETWORK

Sustainability will always remain an issue for an academically based data network and there is no simple answer to the question of how to sustain such a network for the long-term. Self-funding from the home institution is not feasible. Other schemes are possible, such as user fees, etc., but these will only account for part of the costs. However, evolving relationships between the customers, that is the user communities, and the federal funding agencies may provide a partial answer. As we write this paper, things are moving rapidly both within our user communities and with the federal funding agencies (Department of Energy and National Science Foundation in particular). These agencies are themselves going through an evolution of thought on the importance of data stewardship and their roles within the overall data life cycle (e.g., NSF's EarthCube initiative). For some agencies, it makes sense that the agency itself should host some of the data generated by their operations. However, for other data, it imperative that they be hosted outside of the agency or the underlying issues of transparency and trust will remain cloudy and debatable. One of the themes of this paper is that there needs to be long-term partnerships between agencies and academically based data sites, in particular data networks. Another underlying theme is that this improved data stewardship has costs associated with it, i.e., irreducible baseline costs. Thus, as agencies continue to increase their efforts to manage and provide access to data generated by projects and activities they fund, a long-term agency-academic partnership will evolve that includes both funding academically based data networks and relying on these networks to provide some of that public access to federally-funded data. Finally, it is important to note that these academically based data networks have to be self-governing for them to work at all, much less be sustainable; whereas federal agencies may provide financial support, it is the user community that must decide what and how things are done.

SUMMARY

Our interaction with the geothermal community and our experience building multiple data systems has provided many insights into both improving our individual data sites and into sustaining a distributed data network. We are fully aware of the importance of working with federal and state agencies on these endeavors and of international collaborations - we are doing all of this.

The future lies in partnering with the federal funding agencies to continue to build systems that: 1) accommodate the needs of individuals, research teams and projects, and commercial enterprises, 2) provide public outreach and education, and 3) help meet the internal needs of the agency. The network must be technically and operationally flexible to mold to the needs of users and each of the nodes on the network; it must not unnecessarily force users or data sites to conform to the data system. All of these aspects present significant problems to building and maintaining a data system, but none of them are new to us and all of it grows from our roots in the user community.

In summary, the future of an academically based network for geothermal data and the geothermal community is bright, provided we continue to operate under the principles of openness and collaboration; and provided, too, that we remain flexible and responsive to the community we serve as it becomes more knowledgeable and involved, as technologies evolve, and as opportunities for sustainability emerge. So long as the academic groups and the interested federal and state agencies are willing to collaborate there are no insurmountable barriers to creating the dynamic system we envision.

EDITOR'S NOTE

This paper was originally published in the 37th Stanford Geothermal Workshop proceedings and reprinted with permission from the Stanford Geothermal Program.

REFERENCES CITED

Atkins, D.E., Droegemeier, K.K., Feldman, S.I., Garcia-Molina, H., Klein, M.L., Messerschmitt, M.L., Messina, P., Ostriker, J.P., and Wright, M.H., 2003, *Revolutionizing Science and Engineering through Cyberinfrastructure: Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure*, 52pp. (http://www.communitytechnology.org/nsf_ci_report/) or (http://www.nsf.gov/publications/pub_summ.jsp?ods_key=cise051203)

Atkins, D.E., Hey, T., Hedstrom, M., eds., 2011, *NSF Advisory Committee for Cyberinfrastrcuture, Task Force on Data and Visualization* 46 pp.; Accessed 12 December, 2011; http://www.nsf.gov/od/oci/taskforces/

Deloitte Consulting, LLP, 2008, *Geothermal Risk Management Strategies*; a report for Department of Energy - Office of Energy Efficiency and Renewable Energy, Geothermal Program, 44p. www1.eere.energy.gov/geothermal/pdfs/geothermal_risk_mitigation.pdf

Entingh, D., 2002, Princeton Energy Resources International, LLC, *Geothermal Studies and Analyses: Report 6A. Status of DOE Geothermal Technical Report Collections*, a report for Department of Energy - National Renewable Energy Laboratory, 36 p. www.perihq.com/documents/Report%206A.pdf

Hey, T., Tansley, S., and Tolle, K., eds., 2009, *The Fourth Paradigm: Data-intensive Scientific Discovery: Microsoft Research*, Redmond, Washington, 252 pp; Accessed: 2011-03-04; http://research.microsoft.com/en-us/UM/redmond/about/collaboration/fourthparadigm/4th_PARADIGM_BOOK_complete_HR.pdf

Interagency Working Group on Digital Data, 2009, *Harnessing the Power of Digital Data for Science and Society: Report of the Interagency Working Group on Digital Data to the Committee on Science of the National Science and Technology Council*, January, 2009, 60 pp.

Nature Editorial, 2009, "Data's shameful neglect", *Nature*, v 426, no. 7261, 10 September 2009, p. 145.

NISO, 2004, *Understanding Metadata*, National Information Standards Organization www.niso.org/publications/press/UnderstandingMetadata.pdf

NRC, 2002, "Geoscience Data and Collections", *National Resources in Peril*, National Research Council, National Academy Press, Washington, DC, 107 p.

NRC, 2009, *Ensuring the Integrity, Accessibility, and Stewardship of Research Data in the Digital Age: Committee on Ensuring the Utility and Integrity of Research Data in a Digital Age*; National Academy of Sciences, National Research Council; ISBN: 978-0-309-13684-6, 188 pages (http://www.nap.edu/catalog/12615.html).

NSB, 2005, *Long Lived Digital Data Collections: Enabling Research and Education in the 21st Century*; National Science Board, publication NSB 05 40; http://www.nsf.gov/pubs/2005/nsb0540/start.htm

OSTP, 2009, "Harnessing the Power of Digital Data for Science and Society":, National Science and Technology Council, *Report of the Interagency Working Group on Digital Data to the Committee on Science of the National Science and Technology Council*, Office Science and Technology Policy; 60 pp. Accessed: 2011-02-27; http://www.nitrd.gov/About/Harnessing_Power_Web.pdf

Uhlir, P.F., et al., 2009, *The Socioeconomic Effects of Public Sector Information on Digital Networks*, National Research Council, 104pp; http://www.nap.edu/catalog/12687.html