Geothermal Data Management and Best Practices

Jeffrey Benegar\textsuperscript{1} and Robert Kennedy\textsuperscript{1}

\textsuperscript{1}Tetra Tech, 14151 Newbrook Drive, Suite 400, Chantilly VA USA 20151

Email: jeff.benegar@tetratech.com, robert.kennedy@tetratech.com

\textbf{Keywords:} Geothermal data management, relational database, risk mitigation, best practices

\textbf{ABSTRACT}

Data management is an activity that is rarely considered during the life cycle of a geothermal project. “Why data management?” is an uncommon question when it comes to geothermal projects. However, the need for data management is critically important because geothermal development projects consist of multiple activities that typically generate significant amounts of data. The type of data associated with a geothermal project spans a broad spectrum of time, formats, and disciplines (geology, geochemistry, geophysics, reservoir engineering, GIS and remote sensing, for example). These data can be disconnected and siloed among individuals and departments, even in mature geothermal jurisdictions. This disconnectedness can lead to wasted resources, duplicated effort and the failure among team members to make meaningful insights with respect to the geothermal resource. Geothermal development consists of multiple phases in which data that is collected during each phase helps reduce the overall project risk. The geothermal exploration phase consists of acknowledged high-risk activities that generate lots of data. The highest risk and cost occur during the drilling of the first well, when resource uncertainty is still high. This high-risk barrier is frequently the biggest obstacle during a geothermal project. Data management helps reduce the risky nature of geothermal development to the lowest achievable levels with good data that informs decisions at every stage of the project. Spreadsheets are not databases—a modern relational database management system is the best practice to securely organize, integrate and analyze data and help mitigate the risk associated with geothermal development projects.

1. Introduction

Obstacles exist for developing geothermal energy projects, attributable to several causes. One key barrier towards exploiting this source of energy is the lack of accessible, reliable and validated geo-science data on geothermal prospects, cross-referenced with other disciplines. There is a significant amount of disconnection and isolation (i.e., “siloes”) between individuals and departments. Therefore, it is crucial that technical professionals not only have a well-organized data management system but also have the necessary “all round” training to improve their knowledge and skills in data management in order to competently carry out their mandate.

A key aspect of geothermal development is gathering, managing and interpreting technical data (e.g., geologic, reservoir, financial) so that it can be integrated to reduce high-risk activities such
as drilling and well-field development, leading to more successful exploration and power development outcomes. It is the experience of the authors that decision-making expertise from subject matter experts (SMEs, e.g., geologists, geophysicists, and reservoir engineers) is not sufficiently linked with each other nor with government officials, policy makers, regulators and the many stakeholders associated with geothermal activities. The step from technical analysis through to commercially driven project planning is dependent on reliable and readily accessible data.

Meaningful evaluation of any geothermal resource is dependent on the consistent acquisition, storage and management of high-quality data relating to the concession so that the data can be interpreted for exploration planning and conceptual model development. Key decisions can be made on resource value, financing, risk sharing, and licensing when based on well-managed data. Data must also be secure, and any system needs to protect the data from uncontrolled changes and unauthorized access. Ultimately, the purpose of data management is to reduce the risky nature of geothermal development to its minimal achievable level with good data practices that informs decisions at the various stages of development.

The best practice is to develop a relational database management system (RDBMS) in order to securely manage geothermal data and mitigate risk during each phase of a geothermal project. For this reason, mere spreadsheets will not do. The data management system should be based on international best practices as well as input from geothermal and database management experts and be designed to fill the principal needs of a government geothermal agency, other government institutions with related accountabilities, or private industry. These institutional needs include: 1) data collection, 2) integration, 3) accessibility, 4) integrity and 5) security (data and systems). The use of a data management system that includes these considerations help support expert decision-making and risk mitigation for exploration and development of a geothermal power project.

2. Principles and Approach

A key aspect of geothermal development is properly gathering, managing and interpreting technical data (e.g., geographic, geologic, financial), as well as maintaining the data’s integrity, so that it can be integrated and evaluated in order to reduce the technical and financial risk associated with drilling, geothermal field development, and subsequent activities (Mawejeje, et. al., 2018). Exploration risk is best mitigated by experienced data interpretation (IGA, 2013). In many cases, field intelligence and decision-making expertise from SMEs is disconnected from other SMEs and team members as well as government officials, policy makers, regulators and the many stakeholders. The planning and execution of successful commercial geothermal projects is critically dependent on reliable, trustworthy, and readily accessible data by all stakeholders. In addition, high quality data that is supported by safe and reliable access tends to improve resource characterization potential and project bankability (Umbara, et. al., 2016). The use of a relational database management system is the best practice to store, access and utilize valuable geothermal data.

2.1 Databases are not “Spreadsheet, but Better”

A database is an organized collection of data, usually stored and accessed electronically on a computer system. A relational database is a specific type of database that allows users to identify and access data in relation to another piece of data in the database. Usually data in relational databases are organized in tables. A relational database management system (RDBMS) is the software used to store, manage, query and retrieve data stored in a relational database. It is important to note that databases are not simply “spreadsheets but better”. A spreadsheet application, like Microsoft Excel, is commonly used to manage geothermal data. Spreadsheets are inherently simple to understand, learn and use. Excel is readily available, used worldwide by
scientists and engineers and can be used offline (i.e., can be used on devices such as laptops, tablets or smartphones that are not connected to a central computer system). By contrast, databases use tables as a means of storing and retrieving data. Tables are organized as columns (fields or attributes) and rows (records). This tabular structure is similar to spreadsheets, but unlike Excel, most databases are relational, that is, the data between tables can be linked and cross-referenced. These relationships are a logical connection between different tables and are typically based on the interaction of data in these tables. Well-defined relationships (i.e., rules, keys) established in the database allow data tables to communicate and share information with one another and facilitates data searching, organization, and security.

Data integrity is a key difference between databases and spreadsheets. Integrity refers to the reliability and trustworthiness of the data throughout its life cycle. For a geothermal project, the life cycle from original data collection to long-term storage and preservation could last decades. Compromised or inaccurate data is of little use to a geothermal enterprise and can pose a significant financial as well as technical risk.

Databases are ideal for sharing and collaboration of information. Simultaneity is the concept that multiple people can be working on a database at the same time. One person can make a change to the database that is visible to everybody instantly. This capability is not possible with spreadsheets. Since multiple people can access and update the database concurrently, sidestepping the coordination problem which goes as N-squared (i.e., Metcalfe’s Law) a relational database is far more efficient than a spreadsheet and the potential for costly errors related to the data is significantly reduced.

2.2 Data, Data, Everywhere

Geothermal data is extraordinarily diverse and wide-ranging depending on the prospect and the level of exploration which the prospect has been subject to. This data can be structured, semi-structured or unstructured. It may come from a fixed location or time-independent or time-dependent data. In general, there are several broad domains of geothermal data, including: 1) legacy reports (e.g., papers, reports, maps, either physical hard copies or soft copies), 2) physiography and geomorphology, 3) geology, geochemistry, and geophysics, 4) imaging and geodesy/geomatics, 5) wells and boreholes, and 6) socioeconomic (e.g., environmental impact assessments).

These different types of data, while not comprehensive for any one geothermal prospect, are likely to characterize more than 90% of the data that will be encountered over the life span of a project (Anderson, 2013). How much data is this? For a long-term geothermal project, millions of data records can be generated. Because databases store information more efficiently, they can handle volumes of information that would be unimaginable in a spreadsheet. Spreadsheets have record limitations, while databases have unlimited capacity. Finding specific data can be time-consuming or difficult. Relational databases use querying tools to overcome these issues. Legacy data (such as old reports, maps and raw data) may contain valuable information that can “fill in the gaps” of knowledge for a particular geothermal prospect. A geothermal data management system can store this legacy information in data tables or provide hyperlinks to the data. This allows relatively easy access to this data while also providing long-term preservation.

The key to successful management of geothermal data is to collect, organize and disseminate this diverse data based on best practices. The goal is to improve the capacity and decision-making capability of in-country stakeholders by providing them with access to high-quality data tailored to their specific interests.

Illustrative example of the value of comprehensive and modern data management for decision support: Many field reports up to a century old in cursive handwriting on miscellaneous paper
exist in government/corporate archives or public libraries in East Africa. Despite archaic form and age, sometimes the content can be absolutely critical to a modern geothermal project. For example, a multi-million-dollar exploratory drilling project by a foreign aid donor would not have been located where it was had the sponsors seen the report of a foreign oil exploration project in the vicinity which had been undertaken over 50 years before. Interpreting the old drilling logs, and cognizance of a 600-meter offset between legacy coordinate systems (UTM based on colonial-era “Arc60”) and modern systems (WGS84), would have made all the difference. The relevant information existed in the form of a legacy report on a library shelf and was known to exist by a few experts in the related oil and gas field. However, it was not widely known of, nor readily accessible to, the key decision-makers of the geothermal project.

2.3 Databases are not “One Size Fits All”

As a matter of general principle and best practice, any successful data management system must be “needs driven” by the end users (bottom up), not by the architects of the system (top down). According to international best practices in data management, one of the first steps in designing a data system is to write a “requirements document” that specifies the overall purpose of the system (“goal”), then defines the data to be captured (“content”), the use to which it will be put (“functions”), and who will be using it (“users”) (Maweije, et.al., 2018). This requirement document leads to a general model of the database. To be effective and “fit for purpose”, the design of any RDBMS must be informed by business rules. Business rules are the operating policies and procedures that govern a particular geothermal enterprise. Thus, a good RDBMS is not generic; it is highly context-specific based on the needs of the users.

For example, in some East Africa countries, the primary document establishing a geothermal concession is the exploration license (granted to a developer/IPP by the appropriate government licensing authority). Therefore, based on the rules by which these countries operate, it makes sense for licensing information and the data tables created from this data to be the highest-ranking data tables in the relational database. Georeferenced boundaries and dates in the license database thus become primary data elements in the structure of the database as a whole, from which all other data elements flow, and to which all other data elements can be traced back. In other jurisdictions (e.g., California), where geothermal regulations originally were based on the oil and gas industry, which predates geothermal development by decades, the primary element in their data management system is the well. All other data in the database is linked to this element. One way is not better the other. These examples show that organization must decide internally how best to manage its data based upon the rules by which it operates.

3. Database Development

3.1 Database Design Phases

As discussed above, the design of a geothermal data management system must be enterprise-specific. However, commonalities exist with respect to the various phases of the database design process for any particular organization. Best practices for the development phases of a geothermal data management system include:

- Requirements analysis, including:
  - Examination of the enterprise (GRD) being modeled,
  - Interviews with management and SMEs to assess current and future needs,
  - Assessment of information/data requirements for the enterprise,
  - Business rules by which the enterprise operates,
  - Skills assessment of enterprise personnel.

- Building the team, including:
Benegar and Kennedy

- Specification of team leaders,
- Assignment of Data Custodians (DC) and deputies for redundancy,
- Other team member assignments,
- Assignment of the Database Administrator (DBA) and deputies for to assure continuity.

- Data modeling, including:
  - Entity-relationship diagramming,
  - Schema (structure) definition,
  - Data types and constraints,
  - Normalization (collecting, organizing, and then decomposing large data sets into smaller, more manageable ones to eliminate redundancy and duplication, including QA/QC (“cleaning”) of data before populating data tables in the database).

- Building the database, including:
  - Building the data tables,
  - Establishing the relationships (“keys”) between tables,
  - Populating the data tables with QA/QCed data.

- Maintaining the database, including:
  - Periodic update of data tables with new data,
  - Long-term preservation of data and records.

3.2 The Life Cycle of Geothermal Data

All data in a system has a life cycle. For a geothermal project, the various phases of a typical data life cycle are: 1) creation (i.e., acquisition from field collection, compilation of existing and legacy data), 2) storage (i.e., valid data is securely accessible and backed up), 3) usage (i.e. data viewing, processing, interpretation, sharing), 4) archival (i.e., long-term data storage and re-use), 5) destruction (i.e., purging). It is important that quality control procedures are in place to ensure data validity and integrity throughout all the life cycle phases of data in a geothermal project. With respect to data destruction, it is the opinion of the authors that geothermal data should be stored in perpetuity, given the low cost of computer data storage today (either in locally-hosted servers or cloud storage), there should be no reason to destroy data for storage/cost reasons unless required by regulations (e.g., certain confidential data is destroyed after a regulated period of time in some geothermal jurisdictions).

3.2 Data Custodians and Team Building

A Data Custodian (DC) is a key member of the data management team. The DC is the SME responsible for identifying data sets within his/her area of expertise for storage in the database. This is the “point” person for that particular subject matter area (e.g., geochemistry, geology, geophysics). The DC is the person best suited to know and understand every form of data that exists for that particular domain (i.e., physical reports and maps, electronic/digital, georeferenced data). The DC is also responsible for the quality of the data (e.g., data cleaning) that ultimately will reside in the RDBMS. To prevent single-point failure, the Primary DC should be backed up by a Deputy. Furthermore, no DC performs QA/QC on their own raw data from the field—instead, the Primary and the Deputy cross-check each other’s data. The maintenance, disposition and long-term preservation of records and data is another responsibility of the Data Custodian team.

One of the most important aspects of any data management system is the human resources component. This is evidenced by appropriate team building. The goal of team building is to provide redundancy at each layer of the of the data management system by designating a Principal and at least one Deputy—two when possible—to provide triple redundancy. Redundancy is a
basic principle of systems engineering that provides continuity of operation and preserves institutional knowledge within a geothermal organization.

It is best practice that geothermal data management team members should consist of: 1) geoscientists (SMEs from relevant fields such as geology, geochemistry, geophysics, well drilling/drilling engineers, hydrogeology and reservoir engineering), 2) information technology (IT)/information and communication technology (ICT) experts; 3) cartography experts (e.g., GIS, cadaster); and 4) programmers (database/structured query language (SQL), web programming (front end interfaces)). Given the importance of data quality and having only valid data in the database, having geoscientists as part of the data management team is not only best practice but absolutely critical. Data cleaning and subsequent data entry should only be done by a trained scientist/engineer. A non-scientific technician may not fully understand the data nor be able to catch data entry errors on a particular geoscientific data set, leading to invalid data being stored in the database.

### 3.3 Cleaning for Data Integrity, Good File Names/Structure and Best Practices

As discussed above, a key benefit of an RDBMS is establishing data integrity. It is important that data be accurate before it is imported permanently into an RDBMS, because it is very difficult to find and correct problematic data after it is in the database. Data integrity ensures that all false information is excluded from the database while all true information is included in the database (Motro, 1989). The presence of incorrect or inconsistent data can affect interpretation, conceptual models, resource models and lead to increased technical and financial risk. One duty of the Data Custodian is to establish validity of the all the data before it is imported into the database, which is why the DC must be a trained scientist, not an IT/ICT or administrative technician. Data cleaning (or error checking) is the process of detecting and correcting corrupt or inaccurate records. Data cleaning is necessary early in the data management process to ensure its validity. Errors to look for before data is entered into the database typically include (Hellerstein, 2008):

- Data entry errors,
- Physically impossible values,
- Missing of juxtaposed values,
- Outliers,
- Typographic and formatting errors (e.g., blanks, spaces).

### 3.4 Databases and GIS – Making the Connection

Much geothermal data is spatial in nature – there is a geographic component to it. Usually this data is stored and used in software such as geographic information systems (GIS) or cadaster portals. GIS is an important tool for interpretation of geoscientific data, especially during the exploration phase of a geothermal project. A cadaster portal is the official registry of licensed geothermal concessions/areas. A well-designed data management system will link geospatial data in GIS with corresponding data records in the RDBMS. Geospatial data and attributes can logically be stored and used in GIS but can also be referenced via joins and relates back to their corresponding data elements in the database, by means of a unique identifier (key). This eliminates the need for all data to be duplicated in both the database and GIS. For example, the geochemistry results from a concession area could number in the thousands or more of analytical results. There is geographic data associated with each result (e.g., X, Y coordinate, datum). There is no need to repeat this data for every chemistry result. Rather, the geographic data associated with these results should be stored only once in a separate locations data table and also linked to its corresponding representation as a layer in GIS.

### 3.5 A General Model for a Geothermal Data Management System
As discussed above, a geothermal data management system must be enterprise-specific, that is, it must be designed and built based upon the needs of the geothermal organization and the rules by which they operate. However, there are commonalties among geothermal data management systems. Figure 1 shows a general model for a geothermal data management system. At the top level is data (primary data from field collection, existing and legacy data). All data is under the control of the data custodians for that particular discipline. Separate databases will exist within the system based upon the disciplines included (in this example, geology, geochemistry, geophysics, wells and licensing). All the databases and tables are linked together with common elements based on the rules of the system. There are linkages with GIS as well within the data management system. Legacy libraries of old reports, maps, and documents can also be linked to the data management system. Interfaces are developed for both local users within the geothermal enterprise and external users with appropriate access and use restrictions.

Figure 1. A general model for a geothermal data management system based on best practices.

The development of a geothermal data management system requires an investment in time, human resources and money. It is the experience of the authors that a system could be built in as little as one year, but two to three years is more realistic. Costs for building a geothermal data management system range from $200,000 to >$1,000,000 (US). The length of time and money needed to build a full geothermal data management system is dependent on the availability of qualified technical staff (e.g., geoscientists, engineers, IT and GIS professionals), amount of funding, commitment of management, and organizational infrastructure (i.e., hardware, software, servers, etc.). The geothermal data management system should be designed and built to last for decades, given the life cycle of data, geothermal power projects, and regulatory requirements.

4. Conclusions

The need for data management is important because geothermal development projects consist of multiple activities that typically generate significant amounts of data. In many geothermal jurisdictions, data is disconnected and siloed among individuals and departments. This disconnectedness can lead to wasted resources, duplicated efforts and the failure among team members to make meaningful insights. Geothermal development consists of multiple phases in which data that is collected during each phase helps reduce the overall project risk. Reducing the
exploratory drilling risk to its minimum practicable level will lead to more successful exploration and better outcomes for developing clean electrical power as this is the highest risk phase in any geothermal project.

The effective evaluation of any geothermal resource critically depends on the acquisition, storage and management of high-quality data so that it can be consistently and accurately interpreted. Key decisions can thus be correctly made about exploration planning, the resource’s value, financing, risk sharing, and licensing. The development and use of a geothermal data management system is the best practice to securely manage exploration data and mitigate risk.

REFERENCES


Hellerstein, J.M. "Quantitative Data Cleaning for Large Databases." United Nations Economic Commission for Europe (UNECE) (February 27, 2008)


