Proceedings World Geothermal Congress 2020+1
Reykjavik, Iceland, April - October 2021

Building Robust Geothermal Data Management Teams in East Africa

Robert Kennedy[1], Jeffrey Benegar[1], Peter Mawejje[2], Jacinta Achieng[2]

[1]Tetra Tech, Suite 100, 1093 Commerce Park Dr., Oak Ridge TN USA 37830

[2]Geothermal Resources Department, Directorate of Geological Survey and Mines, Ministry of Energy and Mineral Development, Plot 21-29 Johnstone Road, P.O. Box 9, Entebbe, Uganda

robert.kennedy@tetratech.com, jeff.benegar@tetratech.com, mawejje.p@gmail.com, jachieng225@gmail.com

Keywords: Risk mitigation, decision support, geothermal data management, relational database, East Africa

ABSTRACT

A great deal of data has been collected over the past century from the numerous geothermal prospects in the countries of the East African Rift. It spans a broad spectrum of time, formats, and disciplines (geology, geochemistry, geophysics, reservoir engineering, GIS and remote sensing). These data tend to be isolated and siloed among individuals and departments; legacy non-machine-readable data tends to be unintegrated with modern digital data. A modern RDBMS is the best way of integrating this vast amount of disparate information in order to correctly make important decisions, such as where to commence exploratory drilling. We have found that geothermal scientists in East Africa must be "all-rounders" able to understand the language of the multiple disciplines and data types involved, correctly interpret their diverse data, and proficient with both modern data management tools and the ICT foundations those tools run on. Staff must operate according to international best practices in order to properly interface with developers and multilateral aid organizations. Furthermore, East Africa is a dynamic business environment with constant personnel churn, which poses a particular challenge for developing geothermal projects that typically take the better part of a decade to come to fruition, much longer than most other forms of renewable energy. Therefore, a robust team that can be successful in this market over the long term must be extensively cross-trained and have multiple redundancies in all key positions. We have learned that the human resource architecture is as important as the technical data management architecture. We would like to share these lessons with our colleagues.

1. INTRODUCTION

The unsustainable energy mix in East Africa greatly hinders economic development, not to mention environmental damage due to deforestation and issues associated with climate change. The desire is to transform from a largely rural economy based on subsistence agriculture to an urban industrial economy. Currently, a significant portion of the East African population relies on biomass for energy (either wood harvested directly from the forest, or charcoal fuel inefficiently produced from that) because they do not have access to electricity nor clean sources of heat for domestic tasks like cooking and cleaning. Development of geothermal resources is one of the key proposed strategies to diversify the energy mix and sustainably facilitate economic transition, through increased electrification of the East African nations (Uganda National Planning Authority, 2007). Developing direct use applications for geothermal energy will boost several already existing industries and create new ones, in areas such as agro-processing, horticulture, tourism and fish farming.

Obstacles exist to developing geothermal energy in East Africa, attributable to several causes. One key barrier towards exploiting this source of energy is the lack of accessible, authoritative, reliable and validated geo-science data on geothermal prospects. There is a significant amount of disconnection and isolation (i.e., "siloed") 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 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 (geologic, reservoir, financial) so that it can be integrated to reduce the risk associated with drilling and well-field development activities, leading to more successful exploration and power development outcomes. It is the experience of the authors that field intelligence and decision-making expertise from subject matter experts (SMEs, e.g., geologists, geophysicists, and reservoir engineers) is not sufficiently linked 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.

The effective and consistent evaluation of any geothermal resource is dependent on the consistent acquisition, storage and management of the high quality data relating to the concession so that the data can be interpreted for exploration planning, and key decisions can be made on resource value, financing, risk sharing, and licensing. 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.

This paper describes the principles and development of a relational database management system (RDBMS) for geothermal systems. This data management system is based on international best practices as specified by geothermal and database management experts, designed to fill the principal needs of a government geothermal agency, other government institutions with related accountabilities, or private industry. These needs include, data integration, accessibility and integrity as well as data and systems security, which support expert decision-making for exploration and development that is 'fit for purpose' for East Africa.

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. Exploration risk is best mitigated by experienced data interpretation (IGA, 2013). In many cases, field intelligence and decision-making expertise from SMEs is not sufficiently linked with 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.

2.1 Geothermal Data Types

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 or time-dependent data. In general, there are several broad domains of geothermal data, including:

Desktop studies (e.g., papers, reports, maps, either physical hard copies or soft copies),

Physiography and geomorphology,

Geology,

Geochemistry,

Geophysics,

Imaging and geodesy/geomatics,

Wells and boreholes,

Socioeconomic (e.g., environmental, ecological and cultural data).

These classes of information, while not comprehensive for any one particular geothermal prospect, are likely to characterize about 90% of the data that will be encountered over the life span of a project (Anderson, 2013). Some data tends to be unstructured, either not codified, or worse, not converted to machine-readable form. The key to successful development of geothermal power is to collect, manage and disseminate this data based on best practices, more widely than now, while simultaneously ensuring that confidential data obtained from developers is treated appropriately. The goal is to improve the capacity and decision-making capability of in-country stakeholders by providing them with access to data tailored to their specific interests.

The proprietary nature of much of the collected data is another challenge, due to (for example) different formats and different interests of different organizations, and general non-transparency of the private sector, especially that dealing with the risky and costly exploration for geothermal resources or oil and gas (O&G), an industry that geothermal energy shares much with. The special nature of geothermal resources, much more akin to O&G production and distinct from the mining of solid materials and minerals, is now recognized by the national mineral map portal of Uganda, which is publicly available (Figure 1).

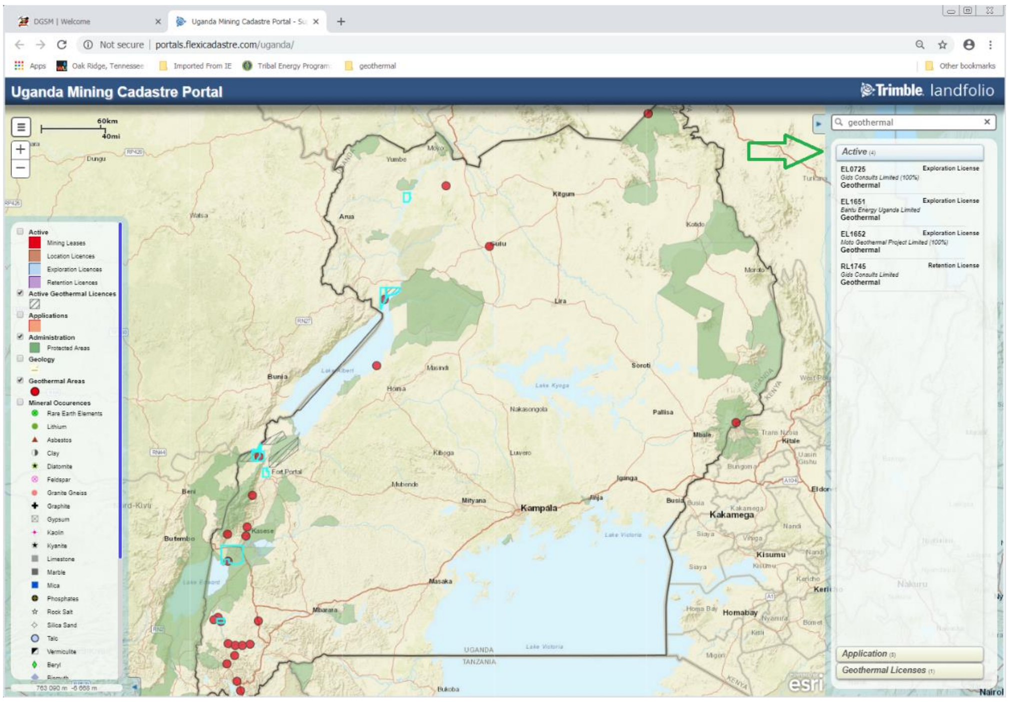

Figure 1: Web portal showing geothermal areas in Uganda

O&G is specifically mentioned herein because the most actionable intelligence to support a decision where to drill may not come directly from geothermal investigations, but rather from a related field such as petroleum exploration.

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 dusty report on a library shelf and was known to exist by a few experts in the related O&G field. However, it was not widely known of, nor readily accessible to, the key decision-makers of the geothermal project.

2.2 Summary of Database Types

These types of databases exist (in chronological order, from simple to complicated):

Flat files, which have existed since at least the 1800s.

Spreadsheets, consisting of simple row and columns (e.g., in late 1800s census records, then with embedded macros and functions when desktop automation began in the 1970s).

Hierarchical DBs, consisting of “parent/child” tables, conceptually similar to e-mail folders. The multiple redundant entries required introduces the danger of inconsistency; also, the hierarchical architecture cannot support “many-to-many” relationships necessary for truly understanding and acting on one’s data.

Network DBs (where “nodes” mean collections of records, “relationships” are equivalent to set structure, and “owner/member” concept superseded parent/child paradigm), came next. Network DB architecture supports “one-to-many” relationships, making them more powerful, but changing a relationship requires changing the set structure.

True relational databases, based on mathematical principles of set theory and predicate calculus, arrived in the 1990s. Specialized terms include: “relations” (means tables), “tuples” (means records) and “attributes” (means fields). The benefits of this approach are:

Each record possesses a unique identifier,

They provide the ability to manipulate and ask questions of the data (i.e., “query”) via structured query language (SQL),

With currently available software, queries can be generated intuitively in a graphical user interface (GUI) without needing programming skills to formulate the query.

After 2000, the “object-relational” model extended the relational DB to object-oriented elements such as classes, encapsulation, and inheritance. Extensions allow an RDBMS to handle storage- and memory-hungry elements objects such as audio, video and architectural/CAD/GIS drawings and images.

2.3 Data Needs and Choice of RDBMS

As a matter of general principle and best practice, any successful information-processing 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”). This requirement document leads to a conceptual model of the database (which should not be confused with the conceptual model of the geothermal resource). To be effective and “fit for purpose”, the design of any RDBMS must be informed by “business rules” – the operating rules that govern a particular enterprise. Thus, a good RDBMS is not generic; it is highly context-specific. At minimum, business rules encompass the enterprise:

Mission, mandate, authority, and limits thereof,

Policies, such as those concerning administrative privileges of various classes of users, and confidentiality,

Procedures (e.g., deadlines, reporting, chain of command).

Example: In Uganda, the primary document instantiating a geothermal concession is the exploration license granted to a developer by the government. Therefore, the license DB becomes the highest-ranking DB. Georeferenced boundaries and dates in the license DB thus become primary data elements in the schema, from which all other data elements flow, and to which all other data elements can be traced back.

The conceptual model in turn informs the decision to choose a relational or non-relational database. The advantages of an RDBMS over the other database types include:

Stored data is logically and physically independent from the applications that create it or uses it. Neither changes to the DB design by users, nor changes by the vendor to DB application(s) will alter or harm the stored data itself.

Guaranteed data consistency and accuracy.

Information retrieval as easy (or hard) as the designer wishes. The data can be viewed in nearly unlimited ways.

Data integrity built in on multiple levels, including at the:

field level to ensure consistency and accuracy,

table level to prevent duplication,

relationship level to assure validity between tables, and

enterprise level to assure data accuracy.

If various geothermal data constructs can be thought of as tools, then the RDBMS is the toolbox they are stored in. For these reasons, the use of a relational database makes the most sense in a geothermal context.

2.4 RDBMS Design Phases

The design of a geothermal RDBMS can be enterprise-specific. Our team proposed the following the design phase for the work being done in East Africa for the Geothermal Resources Department (GRD) under Uganda's Ministry of Energy and Mineral Development (MEMD):

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 GRD,

Skills assessment of enterprise personnel.

Building the team, including:

Specification of team leaders,

Assignment of Data Custodians (DCs) and deputies for redundancy,

Other key team assignments,

Designation of the Database Administrator (DBA) and deputies for redundancy.

Data modelling, including:

Entity-relationship diagramming,

Schema definition,

Data constraints.

Normalization to collect, organize and then decompose large tables and data sets into smaller, more manageable ones to eliminate redundancy and duplication, including the QA/QC ("cleaning") of data before allowing its inclusion in the DB.

3. DATABASE DEVELOPMENT

3.1 Implementation Plan

An implementation plan was proposed and accepted for the development of an RDBMS for GRD that included a web-enabled front-end coupled with the necessary skills development for GRD to build, populate and maintain the system through time. The implementation plan (Figure 2) was a two-year program. The first year focused on skills development for the five key areas of RDBMS development:

1. Security
2. Operating system
3. Webserver
4. Database management
5. Web coding that includes both programming languages and content management systems (CMS)

These five elements, or layers, represent a technology stack ("tech stack"), that is, a development framework that includes all software subsystems to develop and run a full web-enabled platform for an enterprise (i.e., a full geothermal data management system, Figure 3).

The goal of this program was to develop the capacity within GRD for these data management activities. At the end of the first year, the GRD has developed a limited but functional database with a web-enabled front end on the GRD website.

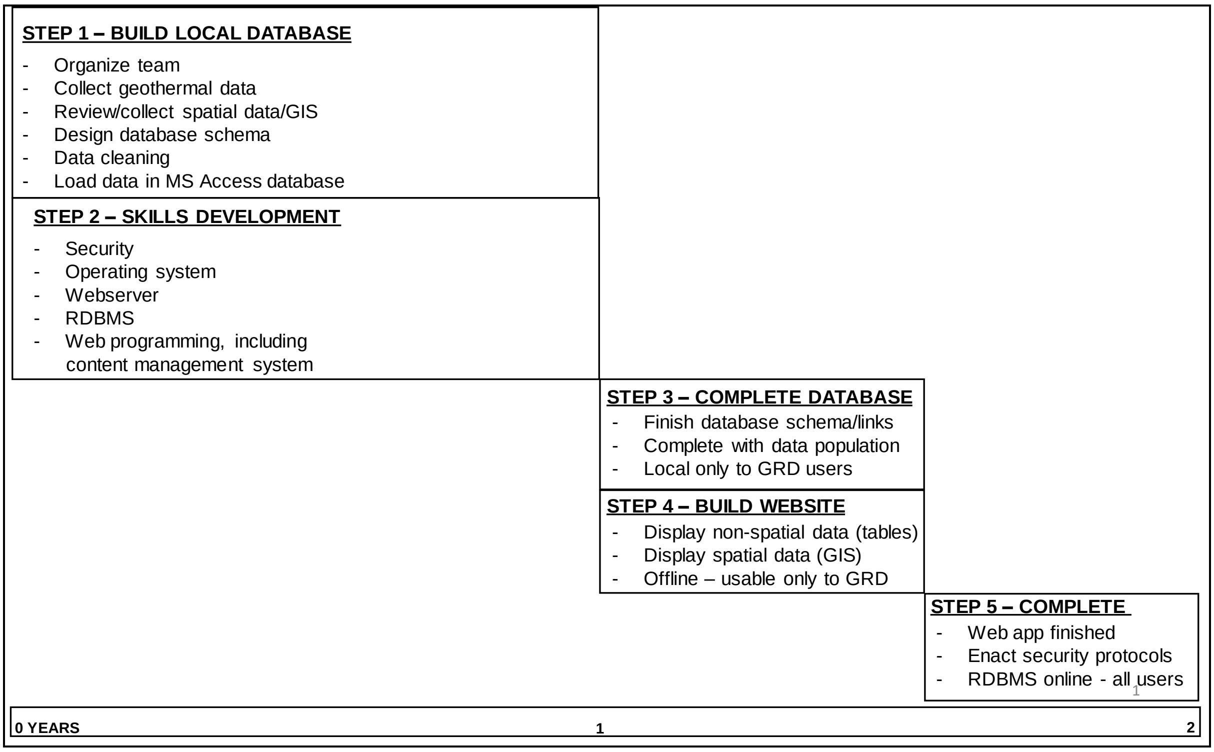

Figure 2: Implementation program for the development of a geothermal RDBMS

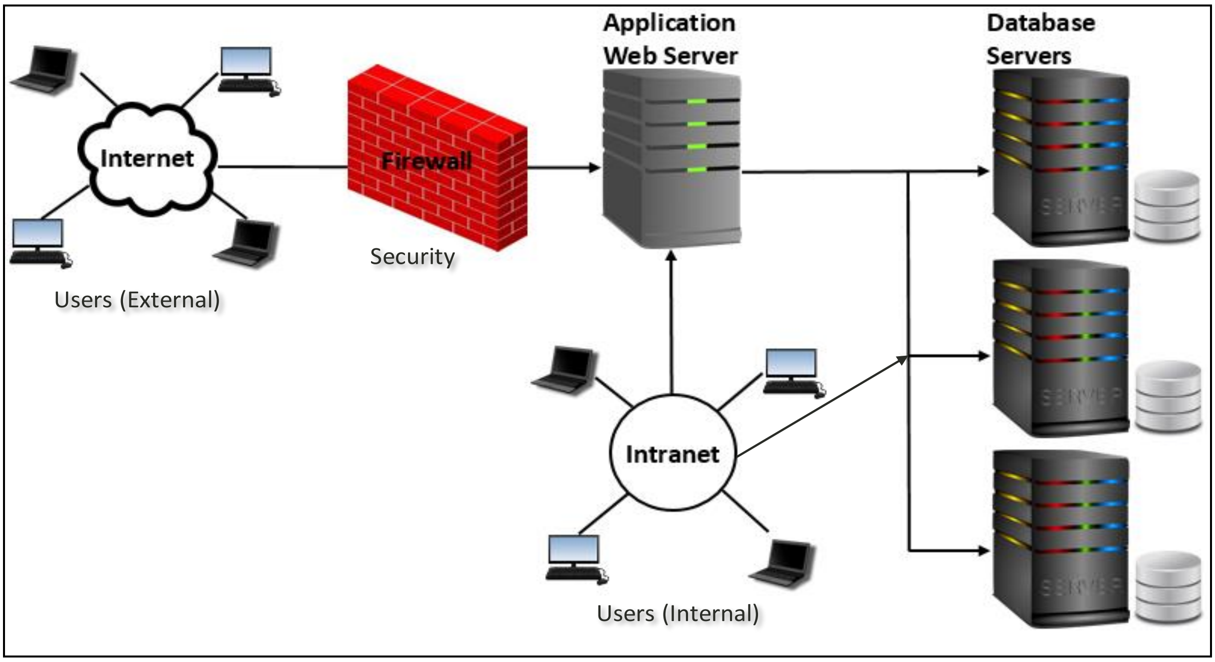

Figure 3: Generic architecture for a geothermal RDBMS

3.2 Data Custodians

A Data Custodian (DC) is the person 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, electronic, paper, georeferenced data, and indices). 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 is 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. An organization chart was developed for each of the five layers of the "tech stack" as well as assigning DCs to each of the five main technical domains that comprise the RDBMS. These domains are:

1. Licensing
2. Geochemistry
3. Geology
4. Geophysics
5. Wells

Licensing is the domain from which all other tables and relationships flow, as shown by the data table t_License below (Table 1) and its accompanying schema depicted in Figure 4.

Field Name	Data Type	Description
License_id	Auto Number	Primary Key
GeoID	Number	GeoID for the license
LicenseName	Short Text	Official name of the license
LicLicenseType_id	Number	Foreign Key, the license type
LicenseAliasName	Short Text	The alias name (if any) of the license
LicenseSignatureDate	Date/Time	Date when the licensed was signed
LicenseEffectiveDate	Date/Time	Effective date of the license
LicenseExpectedExpiryDate	Date/Time	Original date of license expiry
LicenseActualExpiryDate	Date/Time	Actual date of license expiry
Comments	Short Text	General comments

Table 1: t_License data table

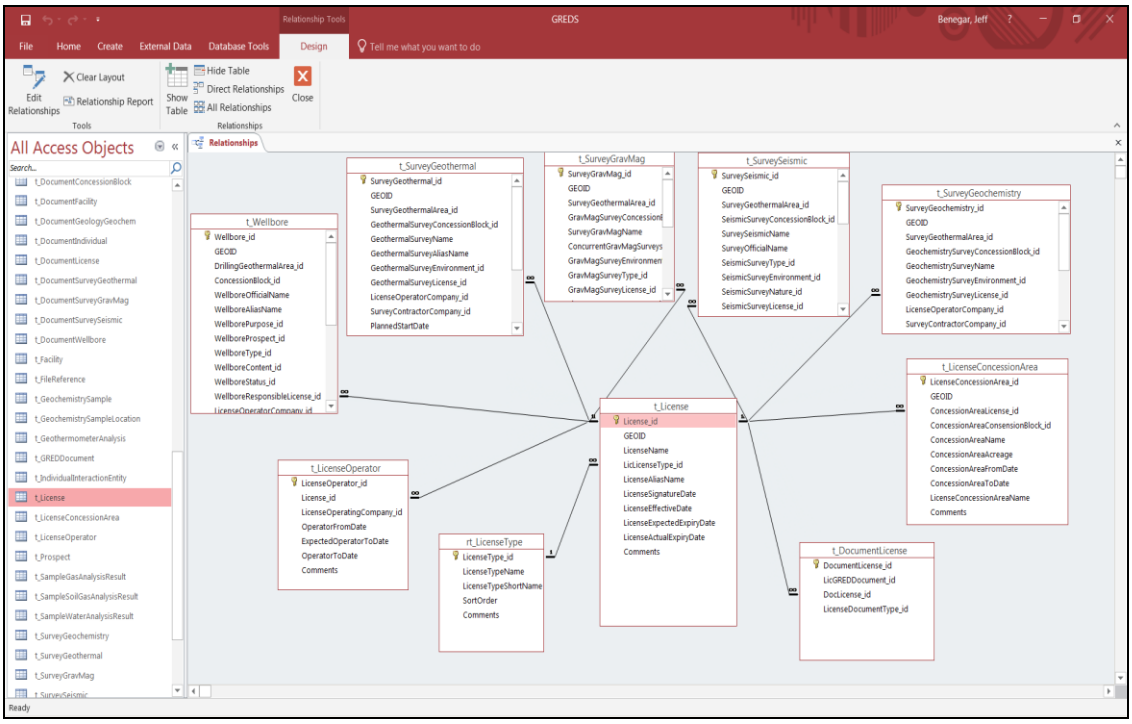

Figure 4: License tables in a geothermal database showing defined relationships between data and reference tables (in this example, this is a Microsoft Access database

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.

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

One of the key benefits of an RDBMS is assuring data integrity. "Integrity" means the maintenance of the accuracy and consistency of data, 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 distort the results of analyses, conceptual models, and resource models. 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 ICT (information and communication technology) technician. Data cleaning (also sometimes referred to as data scrubbing) is the process of detecting and correcting corrupt or inaccurate records from a set of records or data tables. Data cleaning is necessary early in the process to ensure its validity. Errors to look for typically include (Hellerstein, 2008):

1. Data entry or keypunch mistakes
2. Physically impossible values
3. Missing or juxtaposed values
4. Outliers
5. Typographic and formatting errors (e.g., blanks, spaces)

After cleaning, the DC needs to ensure that the data is consistent with other similar data sets in the DB. Because the RDBMS should be useful far into the future, its structure should be logical, navigable, and discoverable—which is an important principle in user interface design. The way the data is organized can be as important as the data itself—thus the DC must consider architecture and provide for posterity. Figure 5 is illustrative of poor practice in assigning filenames and structuring directories/folders. In contrast, Figure 6 depicts the consistent data structuring now being applied at the Uganda GRD.

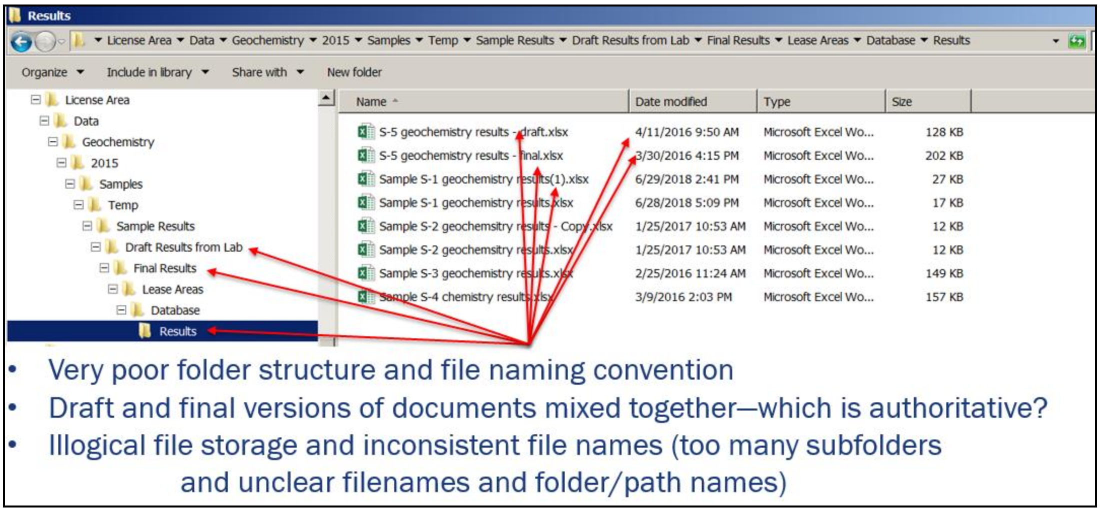

Figure 5: Poor file organization/structure practices

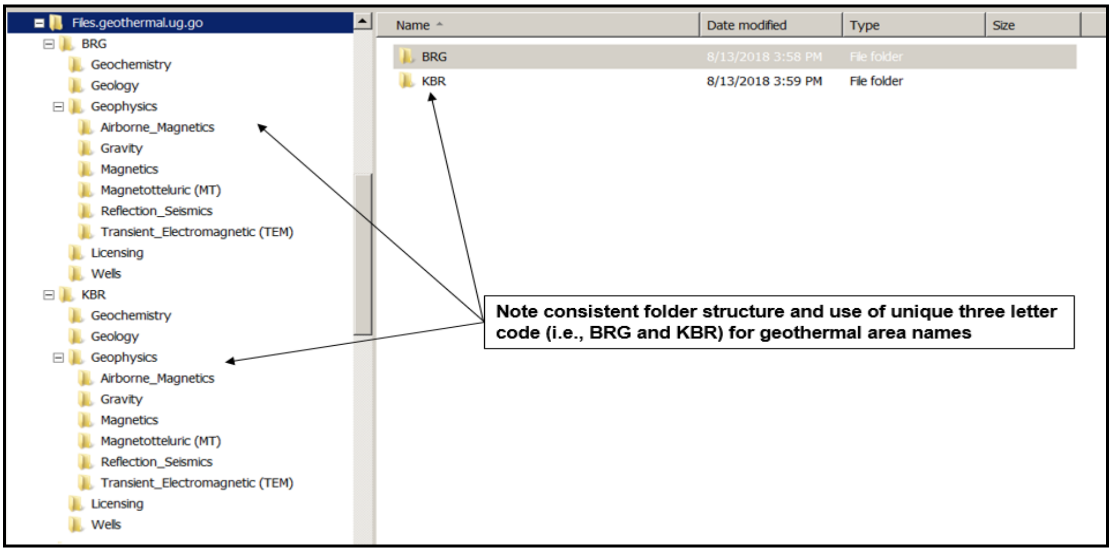

Figure 6: Logical consistent file organization/structure per international best practice assures ready access for posterity

3.4 Geothermal RDBMS for Uganda GRD

The GRD geoscientists developed the schema for the geothermal RDBMS. The RDBMS was given the name "GREDS" (i.e., Geothermal REsources Department Database System). GREDS is a Microsoft Access database that currently contains a number of both data tables and reference tables (i.e., look-up tables). Each table has a primary key defined, as well as foreign keys where appropriate. One major goal of good database design is to remove data redundancy (i.e., duplicate data), because duplicating an entry not only wastes space but worse, needlessly increases the opportunity for mistakes such as keypunch error, resulting in inconsistency, and potentially even failure. To achieve this goal, data was divided into many different subject-based tables such that each piece of information was represented only once. This divided information can then be brought back together by placing common fields in tables that are related. These common fields in different tables are linked together by relationships.

In developing the structure of the database, it became clear to the team that the geothermal license granted to concessionaires by Uganda is the principal "information carrier" via which most other data elements (e.g., geoscientific data) are linked and thus necessarily should be the foundation of the schema. This is the fundamental basis of the GREDS; all data related to each particular geothermal license is linked together via the license identifier (i.e., geochemistry, geology, geophysics, and wells/borehole data). An example is shown in Table 2 below and in Figure 7.

Field Name	Data Type	Description
LicenseConcessionArea_id	Auto Number	Primary Key
GeoID	Number	GeoID for the license concession area
ConcessionAreaLicense_id	Number	Foreign Key, license name
ConcessionAreaConcessionBlock_id	Number	Foreign Key, name of the contract area/concession block that was awarded
ConcessionAreaName	Short Text	Name of the current concession area
ConcessionAreaAcreage	Number	Concession area acreage in sq. km (licensed block area)
ConcessionAreaFromDate	Date/Time	Date when the concession area was created
ConcessionAreaToDate	Date/Time	Date when the concession area expired or was no longer valid (e.g., after relinquishment)
LicenseConcessionAreaName	Short Text	Unique identifier name for the licensed concession area (auto filled by query)
Comments	Short Text	General comments

Table 2: t_LicenseConcessionArea data table

rt_GeothermalArea

GeothermalArea_id	UniqueGeothermalArea_id	GeothermalArea	UTMZone	Topographic	Comment
1	KAG	Kagamba	35M	93/4	
2	KAR	Karungu	35M	93/2	
3	BUB	Bubale	35M	93/2	
4	RUB	Rubaare	36M	85/3	
5	KIT	Kitagata	36N	85/1	
6	IHI	Ihimbo	35M	84/2	
7	KAN	Kanyinabarongo	36N	84/2	
8	BIR	Birara	36N	84/4	
9	RBB	Rubabo1	36N	84/4	
10	KIR	Kiruruma	36N	84/1,2	
11	KIS	Kisiizi	36N	84/4	
12	MIN	Minera	36N	85/3	
13	MUH	Kabuga (Muhokya)	36N	66/3	
14	KIB	Kibenge	36N	66/3	
15	RWG	Rwagimba	36N	66/1	
16	KNG	Kanangorok	36N	9/2	
17	KAI	Kaitabosi	36N	55/3	
18	PAN	Panyimur	36N	29/0	
19	LUS	Lusonga-BH	36N	65/4	
20	AMP	Amuru (Pakele)	36N	13/2	
21	AMR	Amuru	36N	22/2	
22	BRG	Buranga	36N	56/1	
23	KAT	Katwe	35M/36N	75/2	
24	KBR	Kibiro	36N	38/4	

Figure 7: Reference table "rt_GeothermalArea" that contains the unique three-letter identifier for 24 geothermal areas in Uganda as well as other supporting information

3.5 Linking a Geothermal RDBMS with GIS

Much geothermal data is spatial in nature (e.g., faults, lineaments, hydrothermal manifestations, discrete locations for geochemistry samples). As necessary, there is a unique numeric identifier for every record within each DB, and links to data within and to other databases as well as to geospatial data in GIS. This obviates the need for all data to be duplicated in both the DB and the geospatial software such as ArcGIS. Rather, geospatial data and attributes can logically be stored and used in ArcGIS, but can also be referenced via joins and relates back to their corresponding data elements in the DB, by means of this unique identifier: in the GREDS database, it is known as GEOID (GEOgraphic IDentifier). This data element is found in many of the GREDS data tables and has a corresponding data element (identical in value) in a geospatial DB that is stored in ArcGIS (Figure 8).

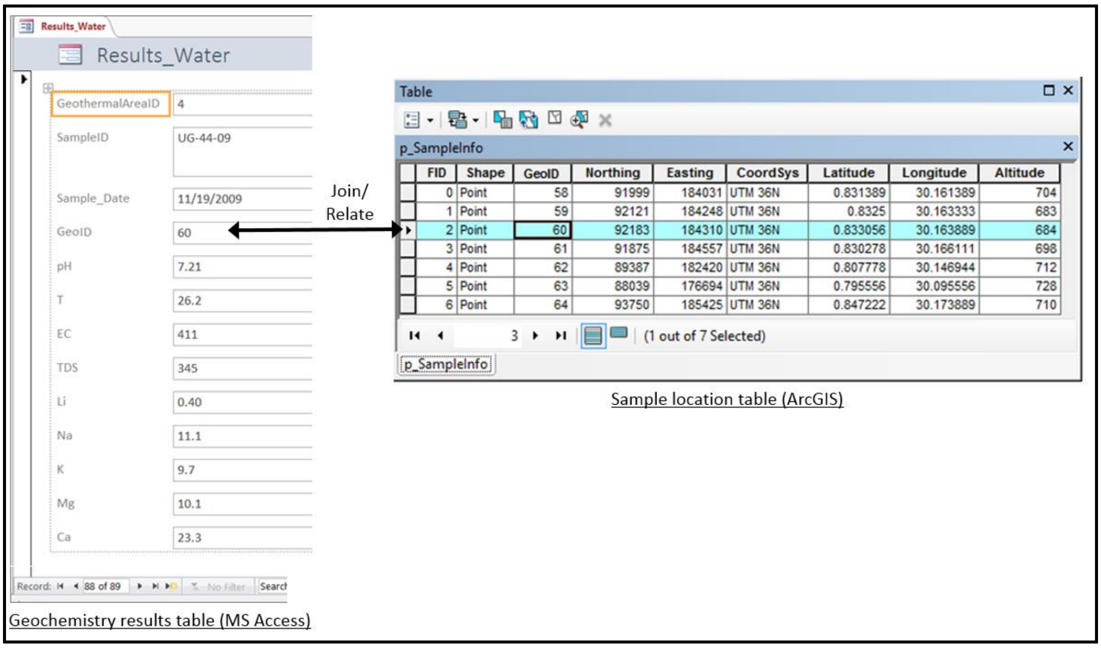

Figure 8: Linkage between RDBMS and geospatial data stored in GIS via the unique GEOID identifier

3.6 Linking a Geothermal RDBMS with Legacy Documents and Reports

A geothermal database management system should provide access to legacy reports and documents. The GREDS database is also being designed to access these legacy reports in Uganda. Various databases exist there which contain unpublished documents, reports and maps that have been scanned and indexed, and are presently accessible via SQL commands. The information contained in these databases represents unstructured data (i.e., soft copy/scanned versions of documents, reports, maps, technical memorandum, etc.). As mentioned in Section 2.1, a well-designed database should be able to access multiple data types (e.g., unstructured, semi-structured and structured data). Providing the links from the GREDS database to these legacy document systems greatly improves the functionality of the database while eliminating the "disconnectedness" of the user (i.e., the geoscientist) from accessing archival or historical information pertaining to a particular geothermal prospect of interest. Figure 9 shows the overall schematic of the GRD database management system.

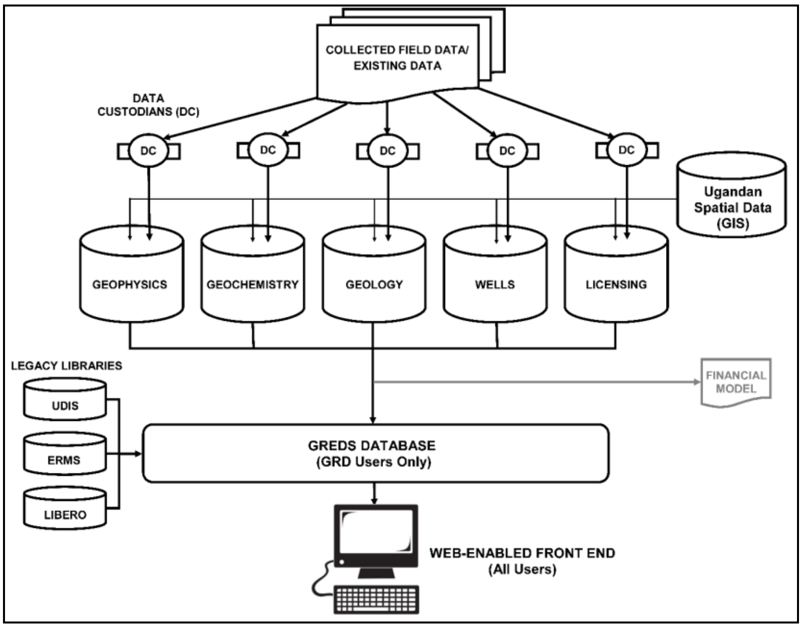

Figure 9: General schematic of the geothermal RDBMS with respect to legacy document and report databases (in this example, for Uganda

4. CONCLUSIONS

A key aspect of geothermal development is gathering, managing, integrating and interpreting technical data (e.g., geologic, reservoir, financial) so that some risks associated with drilling and well-field development activities can be mitigated. Exploration risk is best reduced by experienced data interpretation. The step from technical analysis through commercially driven project planning is dependent on reliable and readily accessible data. Reducing the exploratory drilling risk to its minimum practicable level will lead to more successful exploration and better outcomes for developing clean electrical power. It is the experience of the team that field intelligence and decision-making expertise from subject matter experts (e.g., geologists, geophysicists, and reservoir engineers) is not presently connected well enough with government officials, policy makers, regulators and the many stakeholders associated with geothermal activities.

Therefore, the effective evaluation of any geothermal resource critically depends on the acquisition, storage and management of the data relating to the concession area in a high quality and manner so that it can be consistently interpreted. Key decisions can thus be correctly made about exploration planning, the resource's value, financing, risk sharing, and licensing. However, there is significant disconnectedness (i.e., "siloing") of data and information among individuals and departments in East African nations. In addition, geothermal data is extraordinarily diverse and wide-ranging, which makes it difficult to simply manage. Geothermal data can be structured, semi-structured or utterly unstructured; it may come from a fixed location in space or time, or be variable, or both.

The development and use of a geothermal data management system based on international best practices provides the integration, accessibility and integrity as has been detailed in this paper. Management according to international best practices also provides information, communications, and systems security. It supports expert decision-making for exploration and development that is 'fit for purpose' both in East Africa and the larger world. In addition, the ability of the data management system to link to geospatial data (i.e., GIS data) and legacy documents, reports and maps enhances the usability by all, suitable for the 21st century.

REFERENCES

Anderson, Errol. "Developing a Universal Geothermal Data Management System." Geothermal Resources Council (GRC) *Transactions*, Vol. 37 (2013)

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

International Geothermal Association (IGA). "Geothermal Exploration Best Practices: A Guide to Resource Data Collection, Analysis, and Presentation for Geothermal Projects." IGA Service GmbH and IFC (2013)

Motro, Amihai. "Integrity = validity + completeness" ACM Transactions on Database Systems (TODS), Volume 14, Issue 4 (December 1989)

National Planning Authority. "Uganda Vision 2040", http://npa.ug/wp-content/themes/npatheme/documents/vision2040.pdf (2007)