Addressing Big Data Challenges with Geophysical Data Archiving, Discoverability, and Reuse – 22414

Vicky Freedman*, Judy Robinson*, Kenneth Ham*, and Jon Thomle*
*Pacific Northwest National Laboratory, Richland, Washington, USA

ABSTRACT

Geophysical methods are increasingly used at environmental sites to inform site characterization and perform real-time monitoring of in situ remedial activities. These data often provide a technical basis for decision-making, creating a need to archive both raw and interpreted data in site-managed databases. However, geophysical data are often not stored in site accessible databases because the large temporal and spatial nature of the data make it challenging to incorporate into existing database structures. Moreover, geophysical data are processed to facilitate interpretation, potentially creating data in multiple formats that, over time, may fall out of synchronization with hardware and software versions used to create and analyze them. Although proprietary formats can be efficient for geophysical data processing, they can also hinder interoperability, reuse, and preservation.

This paper discusses the challenges and requirements of archiving large spatial and temporal geophysical datasets, using geophysical surveys executed at the U.S. Department of Energy (DOE) Hanford Site as examples for data archiving. Data access and documentation will be needed to meet regulatory and management needs in support of remedy decisions as the site transitions from initial characterization to implementation and closure. Proper control of the raw and processed geophysical data and associated documentation is essential to their shared, long-term use. The goal is to standardize data archives, capturing pertinent metadata as datasets progress from collection through analysis. The metadata approach identified for geophysical data archiving presented in this paper represents a first step in the standardization of geophysical data archives at Hanford, establishing the needed traceability and data sharing to support site remedy decisions.

INTRODUCTION

Objectives of subsurface characterization include providing information on subsurface structures, contaminant distributions, monitoring of contaminant transport, and delivery of remedial amendments. Borehole-based and well-sampling methods provide discrete point or 1D information and therefore are limited in the ability to spatiotemporally characterize and monitor the subsurface. Geophysical methods can provide 2D or 3D information that can be integrated with conventional well and borehole data to improve spatiotemporal subsurface descriptions.

The Hanford Site has used geophysical methods to complement existing characterization and monitoring methods, yet suitable methods for geophysical data management and archiving do not yet exist. As a result, Hanford geophysical data reside within individual project records. This lack of a shared approach limits the reuse and application of the data and does not meet the need to preserve data after sites reach closure.

While project teams may apply project-specific practices that facilitate wider use, the lack of standardization makes it difficult to maintain those practices beyond the scope and timeline of the project. When data are not placed into a shared repository, they are much less likely to be maintained over time. Data access and documentation will be needed to meet regulatory and management needs in support of remedy decisions as the site transitions from initial characterization to implementation and closure. The reasons that project-specific approaches are not conducive to meeting these needs are provided below:

1. Data are difficult to find for those outside the project.
2. No standard location or method exists for storing and accessing the data.

3. Documentation is directed toward a narrow set of applications and users.
4. Data processing may be narrowly focused to serve specific project objectives.
5. A non-standard vocabulary describing the data may be used.
6. Data provenance may be difficult to ascertain.

A standardized, site-managed approach is needed to meet data archiving requirements for several reasons. First, data reuse avoids unnecessary and redundant data collection. For example, if previously collected data can address a new decision, it should be made available to do so. Second, individual project timelines do not support the longer-term Hanford Site management objectives, which have evolved and shifted depending on new information over a period of decades. The data lifecycle needs to extend beyond the cleanup mission and support the transition to long-term stewardship. A data lifecycle on the order of decades cannot rely on past projects, staff, or existing infrastructure to overcome inadequate metadata, archiving processes, or access control.

DATASET DESCRIPTION

Unlike many other types of Hanford Site datasets, large geophysical data are not easily incorporated into tools such as spreadsheets or databases for ease of access and documentation. Large and complex datasets also need to accommodate a wide variety of formats, sizes, and complexities. Tera- and peta-scale storage may eventually become the norm for geophysical datasets, along with support for multiple proprietary and open data formats. Datasets can also quickly fall out of synchronization with hardware and software versions used in data collection and processing. Although proprietary formats can be efficient for the current processing of geophysical data, they can also hinder interoperability and preservation.

Key examples of large geophysical datasets collected on the Hanford Site are those derived from methods such as electrical resistivity tomography (ERT), electromagnetics (EM), and ground penetrating radar (GPR). These methods can collect spatial and temporal data that can generate large datasets within even modest monitoring periods.

- In 2013, ERT was used to identify flow paths for river water inundation under a process pond loated in the 300 Area. A 352-electrode ERT array was installed to collect time-lapse data [1]. Each data file included 40,454 measurements collected six times a day for more than 80 days, equating to about 2.3 GB of raw data.
- In 2006, ERT data were collected at the Hanford Site B-Complex to image the subsurface distribution of electrically conductive vadose zone contamination [2, 3]. This dataset included 208,411 measurements over 4,859 electrodes equating to about 25.4 GB of data.
- In 2008, an airborne frequency domain EM survey was performed in the 600 Area over the 200-PO-1 operable unit to identify preferential flow paths for groundwater, such as paleochannels and fault zones. Data were collected every 0.2 seconds at six frequencies over 412 linear kilometers [4]. A subsequent airborne time-domain EM survey with a footprint that expanded beyond the 2008 frequency-domain EM survey was collected in the same vicinity, covering approximately an additional 925 linear kilometers using 20 time windows at a rate of four samples/second to produce about 1 GB of raw data.
- Starting in December of 2010, cross-borehole GPR datasets were collected at the BC Crib during the desiccation treatability test. Initially, these data were collected over 30 borehole pairs but this was later reduced to 10 pairs. Each cross-borehole pair had 1360 measurement steps and each measurement step contained 320 time voltage data points. Initial datasets using 30 borehole pairs produced about 2 GB of data. Once reduced to 10 pairs, 0.7 GB of data was collected per set. These data were used, along with neutron probe, ERT, head dissipation sensors, and thermistors,

to spatially identify areas desiccated during this treatability test and to assess the performance of this method [5].

The above examples are of raw Hanford Site geophysical data, but the processed data that is used to make decisions can potentially be larger. As in the last example above, it’s common for ancillary data to be collected alongside geophysical datasets to facilitate ground-truthing or interpretation. For example, ERT data are interpreted with geographic coordinates of each electrode, or if monitoring a tracer injection, the geoelectrical images would need to be validated using data collected regarding the concentration and volume of the injection. Project records may store these files in the same location, but they may be stored without a standard approach for documenting the links between ancillary and geophysical datasets.

GEOPHYSICAL DATA MANAGEMENT

The key to properly managing large geophysical datasets is to focus on the stages through which data progresses from a raw state to processed information and is developed into knowledge. This provides traceability and knowledge sharing at different stages within a data cycle where, for example, proprietary formats could impact sharing and reuse. The stages that data move through are described below and shown in Fig. 1. Similar workflows exist for other geophysical data, including GPR, EM, and seismic methods, which share the need to record spatial and temporal data, but differ in the details needed to describe the surveys and processing approaches.

Geophysical data processing workflows produce several types of files that document the transition from data to interpretation:

- **Raw**. Dataset as collected in an ASCII or proprietary data format from the instrumentation.
- **Processed**. The result of transforming a raw dataset into relevant metrics.
- **Cleaned**. The result of removing anomalous data from processed datasets, leaving only data suitable for its intended use.
- **Analyzed/visualized**. Curated, often objective-specific, visualizations or analyses based on cleaned data.
- **Deliverable**. A formal report or deliverable incorporating an analyzed or visualized output and descriptive interpretation based on the data.

Currently, the only other analogous archival system at Hanford is a file storage system associated with groundwater modeling. Groundwater modeling is another activity that relies on several types of inputs that require documentation and archiving. A structured file storage system is currently used at Hanford to document the groundwater modeling process, but it lacks metadata to make the system easily searchable. This type of system emphasizes the importance of reports that are archived with Hanford data. The linkage to documents associated with geophysical data and their interpretation is a core organizing element for archiving geophysical data at Hanford.

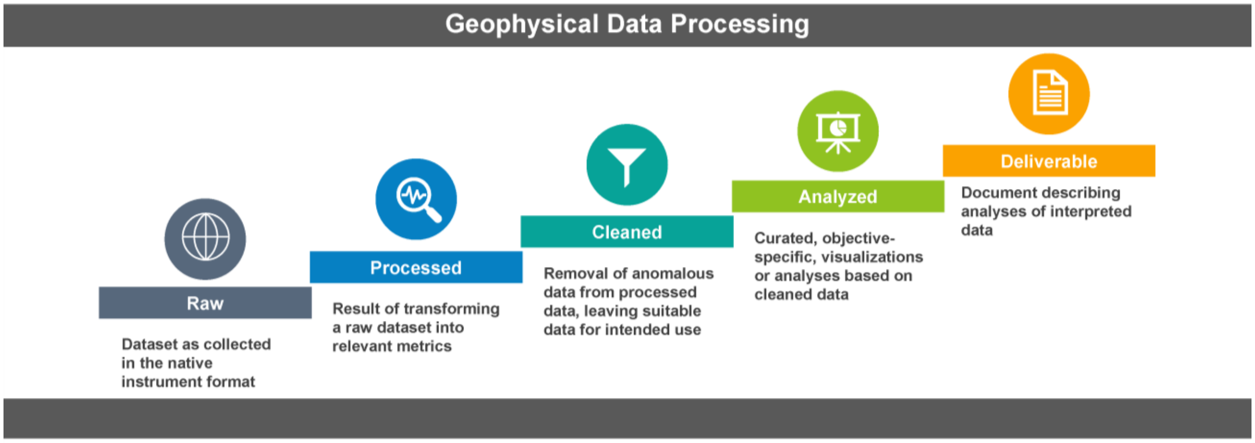
Geophysical Data Processing
Raw
Dataset as collected in the native instrument format
Processed
Result of transforming a raw dataset into relevant metrics
Cleaned
Removal of anomalous data from processed data, leaving suitable data for intended use
Analyzed
Curated, objective-specific, visualizations or analyses based on cleaned data
Deliverable
Document describing analyses of interpreted data

Fig. 1. Example workflow for ERT geophysical data processing.

METADATA

Descriptions of geophysical data and their transformations can be computationally complex, with multiple processing steps, making it difficult for others to discover, retrieve, interpret, and reuse. To this end, metadata – "data about data" – are used to describe the dataset format and content, the circumstances of data collection, procedures used to manipulate or model data, custodianship, data quality, and preservation information.

Metadata are defined as descriptive information about access, use, preservation, interoperability, and management purposes. Metadata can be digital, descriptive text, human- or instrument-generated, and/or managed independently or as part of an information resource. Metadata can be structured or free-text, static (e.g., creator, date of creation, format) or dynamic (as the data experience reuse). General categories of metadata categories are shown in TABLE I, including administrative, descriptive, preservation, system, and technical. For geophysical data, metadata in the descriptive category provide information on the objectives of the data collection and the conditions under which the raw data were collected. The technical category would contain information on the processed, cleaned, and analyzed data, as well as the deliverable. The administrative category may contain contact information on both the raw data and deliverable. The remaining categories are related to the data archive system storing the data.

Although there is no single standard approach for documenting geophysical data, federal agencies are mandated by Executive Order 12906 [6] to use metadata standards endorsed by the Federal Geographic Data Committee (FGDC), including

- Content Standard for Digital Geospatial Metadata (CSDGM) or its extensions for biological data (Biological Data Profile) and shoreline data [7]
- International Organization for Standardization (ISO) series of standards (19115, 19115-2, 19139, etc.) [8]

Although federal agencies, such as the U.S. Geological Survey (USGS), have implemented the FGDC-CSDGM standard for their geophysical data, there are important differences between their approach (focusing on the data collection) and Hanford needs. First, the USGS approach is centered on raw data, rather than on the interpreted data that would be used for decision-making at Hanford. Second, the details of geophysical data collection and processing are largely captured as free-text, which is unconstrained, making their archive less standardized and discoverable than a structured approach. Third, the USGS approach does not capture all stages of the geophysical data workflow from raw data collection and processing to final results and documentation. Finally, Hanford data needs differ from the USGS approach in that additional structure is required to enable discoverability through a geospatial mapping system.

Both FGDC-CSDGM and ISO standards require metadata to be formatted in Extensible Markup Language (XML), with an available stylesheet to make the XML easier to read. Initially, adherence to the FGDC-CSDGM metadata standards for Hanford data archiving was assumed, as it is the current standard supported by the Hanford Site. However, given the potential extensions to the standard needed to describe geophysical data, ISO metadata standards will also be considered because the Hanford Site will eventually migrate to the international standard. To this end, an XML schema will be created that conforms to both ISO and FGDC-CSDGM metadata standards, covering data gaps with a user interface that requires specific input consistent with a Hanford-specific data vocabulary, but results in discoverable free-text within the XML. With this approach, metadata can include sufficient structure to support Hanford requirements while maintaining compatibility with other systems that may use the metadata.

TABLE I. Metadata categories.

Metadata Category	Description	Example(s)
Administrative	Management of data and other resources	Contact information for use; rights and responsibilities of use (e.g., raw data and deliverable)
Descriptive	Description, identification, and context	Discipline-specific tags to identify context under which data were created, including geospatial and temporal information, methodology, protocols, and other scientific descriptors; can also include metadata that would be useful for non-expert users (e.g., raw data)
Preservation	Description of preservation measures used to maintain data	Documentation of condition and steps taken to preserve data, such as migration of data to other systems
System	Description of system and metadata behavior	Metadata standards, hardware/software requirements, networking, security protocols
Technical	Domain-specific description	Data dictionary of geophysical data collection, processing, and interpretation (e.g., processed, cleaned, and analyzed data)

Geophysical dataset discovery through the Hanford Maps (HMAPS) system is a site-specific requirement for a geophysical data archival system at Hanford. In HMAPS, an area can be drawn on a site map to query for available information including raw and interpreted data. To demonstrate the future HMAPS-like functionality for geophysical data, 2D ERT images have been integrated into the ORIGEN module of SOCRATES (socrates.pnnl.gov) [9]. Fig. 2 shows a screenshot of a selected survey, depicting the visualization of the 2D planes within the 3D geological framework model, with a 2D comparison between the ERT image and horizontal lines demarcating the individual geological units in the bottom half of the frame.

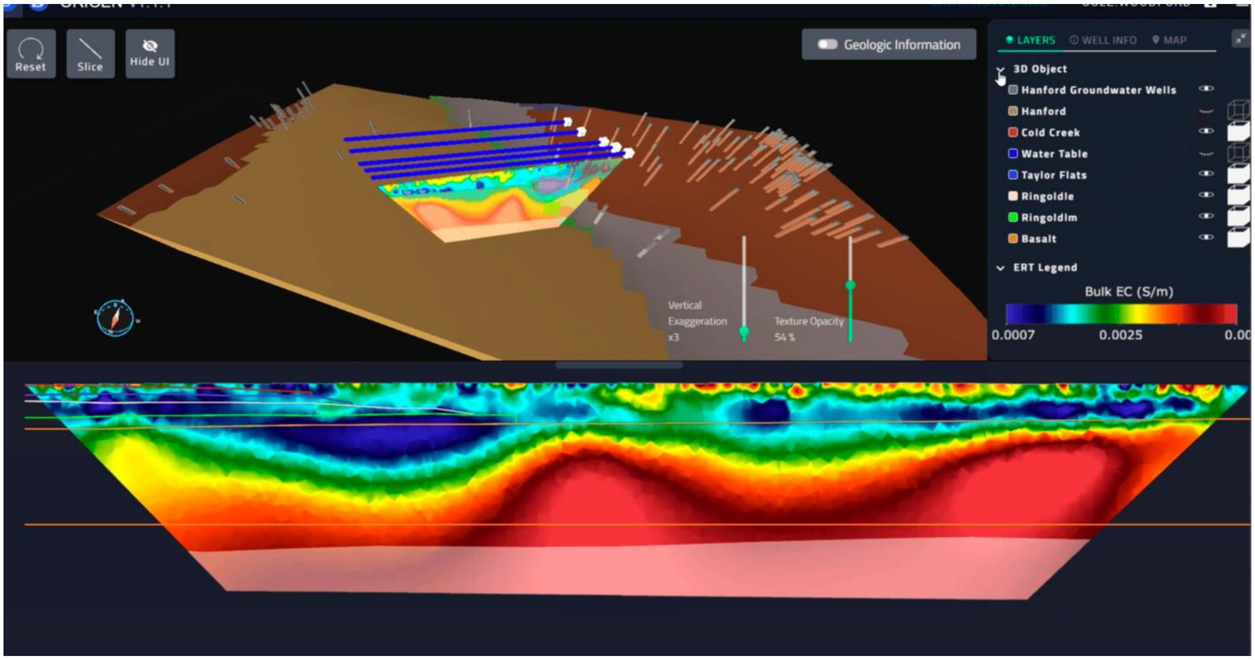

Fig. 2. Demonstration of geophysical data integration into a mapping tool.

CONCLUSIONS

Hanford will need to adopt a framework for geophysical data archiving that accommodates large volumes of heterogenous data in accessible repositories. Two key aspects of data repository accessibility include a user-friendly interface and efficient process for capturing data, as well as a data querying interface that can readily discover geophysical survey data through metadata and HMAPS-like geospatial searching. In addition to metadata, geophysical data archiving will require a data management plan that identifies a framework for integrating data storage, retrieval, and access systems. Initial integration of geophysical data will leverage existing systems at Hanford for data archiving to make data storage and retrieval straightforward, configurable, and sustainable. Policy and procedures will also be established within the existing framework, including policies for data use and restrictions, permissions for contributors to the data repository, and specifications for data formats and metadata. The implementation of these data requirements will make it possible to achieve site goals associated with the archiving and management of large geophysical data sets, and allow for the reuse of datasets to explore new questions, thereby avoiding the loss of data access once projects end. The approach presented in this paper represents a first step in the standardization of geophysical data archiving at Hanford, with potential applicability to other sites.

REFERENCES

1. T. C. JOHNSON, R. VERSTEEG, J. N. THOMLE, G. E. HAMMOND, X. CHEN, and J. M. ZACHARA, “Four-dimensional electrical conductivity monitoring of stage-driven river water intrusion: Accounting for water table effects using a transient mesh boundary and conditional inversion constraints,” *Water Resources Research* 51, 8, 6177-6196 (2015). doi:10.1002/2014WR016129
2. D. RUCKER, M. LEVITT, G. O. BRIEN, and C. HENDERSON. S*urface Geophysical Exploration of B, BX, and BY Tank Farms at the Hanford Site: Results of Background Characterization with Ground Penetrating Radar*. CH2M Hill Hanford Group, RPP-34674, Rev.0, Richland, Washington (2007).

3. T. C. JOHNSON and D. M. WELLMAN, *Re-Inversion of Surface Electrical Resistivity Tomography Data from the Hanford Site B-Complex*, PNNL-22520, Pacific Northwest National Laboratory, Richland, Washington (2013).
4. DEPARTMENT OF ENERGY, *Interpretation of Airborne Electromagnetic and Magnetic Data in the 600 Area*, SGW-47839, Rev. 0, Prepared for the U.S. Department of Energy by CH2M Hill Plateau Remediation Company, Richland, Washington (2010).
5. T. C. JOHNSON, W. J. GREENWOOD, C. E. STRICKLAND, M. J. TRUEX, V. L. FREEDMAN, G. B. CHRONISTER, and D. F. RUCKER, "3D Characterization and Time-Lapse Imaging of the Desiccation Treatability Test at the Hanford BC-Cribs and Trenches Site using High Performance Electrical Resistivity Imaging," Waste Management 2012, February 26 - March 1, Phoenix, Arizona (2012).
6. FEDERAL GEOGRAPHIC DATA COMMITTEE. *Content Standard for Digital Geospatial Metadata*, Reston, Virginia (1998). https://www.fgdc.gov/policyandplanning/executive_order
7. FEDERAL GEOGRAPHIC DATA COMMITTEE. *Content Standard for Digital Geospatial Metadata*, Reston, Virginia (1998). https://www.fgdc.gov/standards/projects/FGDC-standards-projects/metadata/base-metadata/v2_0698.pdf.
8. INTERNATIONAL ORGANIZATION FOR STANDARDIZATION. Geographic Information – Metadata (2003). http://www.iso.org/iso/catalogue_detail.htm?csnumber=26020
9. SOCRATES (Suite Of Comprehensive Rapid Analysis Tools for Environmental Sites). Pacific Northwest National Laboratory. http://socrates.pnnl.gov, Accessed September 25, 2021.

ACKNOWLEDGEMENTS

This document was prepared by the Deep Vadose Zone - Applied Field Research Initiative at Pacific Northwest National Laboratory. Funding for this work was provided by the U.S. Department of Energy (DOE) Richland Operations Office. The Pacific Northwest National Laboratory is operated by Battelle Memorial Institute for the DOE under Contract DE-AC05-76RL01830.