The Data Dictionary: A View Into the CTBT Knowledge Base

Ellen R. Shepherd, Ralph G. Keyser, Hillary M. Armstrong, Eric P. Chael, Christopher J. Young

Sandia National Laboratories

Sponsored by U.S. Department of Energy
Comprehensive Test Ban Treaty Research and Development Program

ABSTRACT

The data dictionary for the Comprehensive Test Ban Treaty (CTBT) knowledge base provides a comprehensive, current catalog of the projected contents of the knowledge base. It is written from a data definition view of the knowledge base and therefore organizes information in a fashion that allows logical storage within the computer. The data dictionary introduces two organization categories of data: the datatype, which is a broad, high-level category of data, and the dataset, which is a specific instance of a datatype. The knowledge base, and thus the data dictionary, consist of a fixed, relatively small number of datatypes, but new datasets are expected to be added on a regular basis.

The data dictionary is a tangible result of the design effort for the knowledge base and is intended to be used by anyone who accesses the knowledge base for any purpose, such as populating the knowledge base with data, or accessing the data for use with automatic data processing (ADP) routines, or browsing through the data for verification purposes. For these two reasons, its is important to discuss the development of the data dictionary as well as to describe its contents to better understand its usefulness; that is the purpose of this paper.

Keywords: knowledge base, data dictionary, datatype, dataset

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy under Contract DE-ACO4-94AL85000.

OBJECTIVES

In support of the Comprehensive Test Ban Treaty Research and Development (CTBT R&D) program, the United States Department of Energy (DOE) is developing a CTBT knowledge base that includes both the content and the framework for storing that content. The knowledge base is intended to improve the ability of the U.S. monitoring community to process and analyze sensor data in support of monitoring the CTBT. All phases of data processing, from automated processing through interactive evaluation of events, are expected to benefit from information contained in the knowledge base.

The data dictionary is a tangible result of the design effort for the CTBT knowledge base and is intended to be used by anyone who accesses the knowledge base for any purpose, such as populating the knowledge base with data, or accessing the data for use with automatic data processing (ADP) routines, or browsing through the data for verification purposes. It is a catalog that describes the data, its use and location, and how to access it for use in the CTBT monitoring efforts.

INTRODUCTION

The design effort for the contents of the CTBT knowledge base project has concentrated over the past year on identifying users of the knowledge base, understanding the framework for its intended use, and characterizing the constituent data components. It quickly became obvious that the volume of data to be stored in the knowledge base is enormous, so one focus of the design efforts became an issue of prioritizing which data to include initially. In addition to the sheer volume of data within the knowledge base, we also identified a large number of formats of data components such as waveforms, gridded information, CSS databases, tessellation grids, tables, and formulas. In this paper, we will summarize the development process and the resulting data dictionary that describes the design of the contents of the knowledge base.

The purpose of the knowledge base is to provide a single, comprehensive source of data required for analysis in support of the CTBT monitoring effort. The analyses may be performed via automated data processing routines such as DFX (signal detection) or GA (event association), or the analyses may be performed interactively by an analyst. Probably one of the most significant contributions of the knowledge base to the CTBT monitoring efforts is the fact that having a single, comprehensive source of data facilitates reproducibility and traceability of analysis results. Another important benefit of the knowledge base is that it is easier to maintain correctness and currency because the data shares a common indexing methodology.

DEVELOPMENT OF DATA DICTIONARY

The data dictionary was produced to reflect the design of the knowledge base; it is intended to make the knowledge base accessible to all users. To better understand the usefulness of the data dictionary, it is necessary to summarize the development of the knowledge base to date, which is accomplished in this section.

Over the past year, we have performed the following tasks with the goal of developing the contents of the CTBT knowledge base:

- *Identified users.* Users consist of anyone who accesses the knowledge base for any purpose, such as populating the knowledge base with data, or accessing the data for use with ADP routines, or browsing through the data for verification purposes. Users may be high level managers, evaluators, analysts, researchers, code developers, for example. Users were asked for input for knowledge base requirements.
- *Identified and characterized automated and interactive applications.* The interface and data requirements were identified, such as those requiring iterative queries or "fast" response time. Some applications are already well defined, while others comprise future functionality. For each application we identified specific data requirements and categories.
- *Identified sources of existing and potential research data.* These include other national laboratories, universities, and commercial sources.
- *Identified relevant metadata for each type of data.* Metadata is data about data, such as date of collection, sensors used, who collected the data, and what type of processing the data underwent. We looked to the Federal Geographic Data Committee (FGDC) for its standard that Federal agencies are instructed to use to document new geospatial data, and ensured that all datasets adhere to this standard. The data dictionary has more details on the FGDC standard and how it applies to the knowledge base contents.

While gathering data requirements for the content of the knowledge base, we also developed requirements and characteristics for the functionality of the knowledge base. These considerations helped focus the design and implementation of the knowledge base.

For example, the main activity of the knowledge base will be data retrievals, with updates being performed relatively infrequently. This influences the selection and implementation of database technology because it puts a higher emphasis on accuracy and speed of retrieval, and less emphasis on database management activities associated with transaction processing such as mirroring and transaction logging. Additionally, we imposed the requirement that the knowledge base had to accommodate different formats of datasets provided by researchers, so that the knowledge base did not introduce artificial methods of collecting data. This requirement also influenced the selection of the underlying database design and management scheme. Further, we ensured compliance with the FGDC metadata standard to facilitate data exchange with geospatial and government entities. It is readily apparent that the size of the metadata alone can match or easily exceed the size of the data itself, therefore, we had to ensure that the database management system could accommodate the volume of metadata.

In addition to the requirements we imposed on the knowledge base, we solicited analysts and other users for requirements for future analytic capability. It quickly became obvious that the volume of data to be stored in the knowledge base combined with all desired functionality is enormous, so then the focus of the design efforts became an issue of prioritizing which data and functions to include initially. We decided to support existing analysis capabilities (data and system interfaces) while planning for future capability.

After identifying all types of data--existing, anticipated, and metadata--we began to look at numerous, various methods of organizing the data for representation in the knowledge base.

Many factors were taken into account in determining an optimal representation such as anticipation of future data types; inclusion of all data types; intuitive appeal to domain experts; exploiting underlying similarities in data across sensor technologies; and independence of database management system implementation.

Many representations were considered:

- traditional relational database schemas
- object oriented representations
- sensor domain representations such as points, stations, and paths
- organization by monitoring technology (for example, all data associated with seismic monitoring or with hydroacoustic monitoring)
- organization by analysis capability (for example, all data associated with event interpretation or all data associated with signal propagation).

Each of these was rejected for various reasons, either because they failed to encompass all types of data, or worse, they failed to take advantage of similarities in the data.

Finally, we derived a representation that organizes information in a fashion that allows logical storage within the computer. We developed two organization categories of data: the datatype and the dataset. The datatype is a broad, high-level category of data, and the dataset is a specific instance of a datatype. The datatype and dataset concepts are described in more detail in the following section, "Contents of the Data Dictionary." The advantage of this representation is that there is a fixed, relatively small number of datatypes, with a potentially large number of *instances* of the datatypes known as datasets. Eventually the bulk of the knowledge base, and thus the data dictionary, will be dataset definitions as new datasets are added.

The concept of datatypes and datasets facilitates development of datasets because the detailed list of defined datatypes provides a common framework for documenting datasets. The defined *datatypes* are of primary interest to *developers* of datasets because of this framework. The *dataset* definitions are of primary interest to *users* of the datasets because the definitions document the contents of the individual datasets, such as how the data was collected and how it should be used.

In addition to developing the design and the structure of the knowledge base, we evaluated and selected database management technology with which to implement the knowledge base:

- Researched underlying database technologies, evaluating for capability to handle multiple data types and custom data types; scalability; speed of retrieval; compatibility with existing platforms; data sets and applications; and technical support.
- Acquired supporting database management technologies and training. Primarily this is the Spatial Database Engine (SDE) technology combined with Oracle Database Management System.

The data dictionary represents the result of all these development stages. Most of these development details are not included explicitly in the data dictionary, but are expounded here to show the stages in designing the knowledge base and thus the data dictionary.

CONTENTS OF THE DATA DICTIONARY

As a catalog, the purpose of the data dictionary is to describe the contents of the knowledge base in context of CTBT monitoring applications and to provide a guide to the location and use of the data. The contents of the knowledge base are represented by datatypes, which are a broad, high-level category of data, and by dataset, which is a specific instance of a datatype. Therefore the data dictionary defines each datatype and then defines specific datasets.

The knowledge base, and thus the data dictionary, consist of a fixed, relatively small number of datatypes. The defined datatypes encompass all existing and foreseen data requirements, and provide the framework for developing and documenting datasets. Datatype definitions are of interest primarily to the developers of datasets because the definitions provide structure for the datasets.

As the knowledge base is populated with new datasets, the data dictionary will become thick with dataset definitions. The dataset definitions are of primary interest to users since the definitions describe the details particular to the dataset such as methods of collection, units of measure, geographical range, and restrictions on usage.

The data types for the knowledge base are defined from the computer's view to take advantage of the similarities in data types across all sensor technologies, and to remove dependencies on automated data processing functionality. An overview of the constituent data types is shown below. Brief descriptions of each data type follow; please see the data dictionary for more details and examples.

Knowledge Base Datatypes

- 1.0 CSS Information
  - 1.1 Reference Events
  - 1.2 Network and Station Information
- 2.0 Calibration Information
  - 2.1 Referenced by Lat, Lon, and Station
    - 2.1.1 Referenced by Lat, Lon, Station, and Phase
      - 2.1.1.1 Basic Grid Information
      - 2.1.1.2 Augmented Grid Information
- 3.0 Contextual Information
  - 3.1 Discrete features
    - 3.1.1 Basic Grid Information
    - 3.1.2 Augmented Grid Information
  - 3.2 Linear features
    - 3.2.1 Basic information for linear features
    - 3.2.2 Augmented information for linear features
  - 3.3 Area features
    - 3.3.1 Basic information for area features
    - 3.3.2 Augmented information for area features
- 4.0 Supporting Datatypes
  - 4.1 Tabular
  - 4.2 Formula (subroutines, algorithms)

CSS Information stands for Center for Seismic Studies Information and is a reference to the CSS 3.0 Database Schema. The CSS schema is widely used within the US monitoring community for operational databases, so there are numerous software applications that understand the CSS schema and can take advantage of data stored in that format. The CSS schema is also likely to be one of the accepted formats for the international CTBT community as well. This schema has been designed for storing event information including the associated arrivals, measurements, and station information. Because of its wide acceptance, common use, and suitability, it was chosen as the storage schema for event information in the knowledge base.

Reference Events are a linked collection of information containing as much detail as needed about a particular event. They are typically well characterized, well-located events of a particular type in an area. Metadata should include the expected use for these events, any known limitations on how they should be applied, and information about the collection and processing (also known as its "lineage") of the event data.

Network and Station Information describes the sensor stations, and provides long-term average information about the station performance and its background noise environment. This is historic or average information as opposed to current or real-time information since the knowledge base is not expected to handle rapidly changing information like the operational database.

Calibration Information has also been called correction data or gridded parameter data. It contains items such as the regional seismic travel-time corrections, hydroacoustic travel times, and other path dependent information. The datatypes defined in this area are largely intended to directly support automated processing applications such as EVLOC, magnitude calculations, etc.

Some of the datatype definitions will refer to **Gridded Information.** This is information associated with points on a grid corresponding to points on the earth, identified by latitude and longitude. The grid may be regular or irregular, depending on the method of collection or generation of data. For intermediate values in the grid, interpolation may be performed, if the provider of the dataset indicates this is permissible.

Basic grids are simple in structure, having only one or a few values associated with each point. **Augmented grids** have additional data structures (tables or an image, for example) at each point. The data structure may be referenced by a pointer. Some of the grids may be associated with a given station, and some of the grids may be referenced by station and phase.

Contextual Information is the broadest of the general information categories in the knowledge base. It contains information useful to understanding the larger context around a particular event, so most of the information in this category is intended for use by someone evaluating an event that needs a more in-depth analysis. One might also think of this category of information as "background" material since it includes everything from coastlines and city locations to depth to Moho maps and ocean temperature profiles.Within this category, there are **Discrete Features,** which are linked to a single, discrete latitude and longitude pair; there are **Linear Features** such as rift zones and faults; and there are **Area Features** such as population centers and salt deposits.

Supporting Datatypes are effectively a sub-category of datatypes, but since it could apply to any of the other datatype categories, it made sense to call it out at this level. These are simply "containers" for complex types of data such as multi-dimensional tables or mathematical algorithms. There are two types of Supporting Datatypes: **Tabular** and **Formula.** Tabular datatype is available for use by other data types as a storage mechanism. Data is stored in tables, which may be two- or three-dimensional. The format for each table is specified in an associated dataset definition, also included in the data dictionary. Formula datatype is a supporting data type, available for reference by other datatypes. These could be executable computer codes or subroutines, interpolated or scripted software, or algorithms with parameters.

Following the datatype definitions in the data dictionary are the dataset definitions. For each dataset definition, there are three sections of information: high level description, conceptual detail, and technical detail. Each section has a specific audience, and provides detail relevant to that audience. For example, the high level description is targeted to users who are not domain experts or do not need implementation detail.

The high level description gives an overview of the dataset and lists points of contact and security information. Then there is a conceptual detail section that details the purpose of the dataset, its time period, lineage (processing steps that the dataset has undergone), and attribute accuracy. Finally, the technical detail section describes data storage description and access information.

There are three types of dataset definitions in the data dictionary: actual, preliminary, and sample. The actual data set definitions describe existing datasets. The preliminary dataset definitions define anticipated datasets; the definition may not yet be complete. A sample dataset is given for the purposes of illustration; the dataset does not really exist and is not anticipated to exist. The heading for each definition will indicate whether it is an actual, preliminary or sample dataset. Datasets that currently exist in the knowledge base are the IASPEI 91 Travel Time Tables and Digital Chart of the World (DCW).

BENEFITS OF THE DATA DICTIONARY

The data dictionary is a single, comprehensive catalog for describing the contents, location, and access methods for the contents of the CTBT knowledge base. It benefits analysts, evaluators, users who browse the knowledge base, developers, researchers who populate the knowledge base with datasets, and high level managers. There are many direct and substantial benefits from using the data dictionary:

- Since it is a single reference for all users, it provides a common framework for all research, development, and user perspectives. This is probably one of the most important benefits.
- The data dictionary helps prioritize research by making it possible to identify areas where more research data is needed.
- The data dictionary is a single source describing the contents of the knowledge base, which will make it easier to maintain currency and historical data.
- It identifies source, currency, and quality of data to aid users in selecting and using datasets.
- It contains a high level description for each datatype and each dataset so that users other than domain experts can understand the contents of the datatype or dataset.
- It describes the format for each datatype and dataset so that developers know what to expect when writing interface software.
- In most cases, the data dictionary does not require specific formats for datasets provided by researchers to avoid introducing artificial methods of collecting data. In the few cases where specific formats are required, such as CSS database formats, these instances and the associated formats are well-documented.

- It defines the requirements for metadata for each datatype and dataset so that (1) researchers will know what to provide when submitting datasets and (2) users will know the source and history of the datasets.
- It has the capability to include hyperlinks to and from browsers to provide direct viewing of the constituent data.

FUTURE DEVELOPMENT OF THE DATA DICTIONARY

Future development of the data dictionary is driven by future development of the knowledge base. As the knowledge base is refined, and as datasets are added to the knowledge base, the data dictionary will be updated to reflect these enhancements.

To help maintain viability of the data dictionary, it will become an on-line document rather than a static paper document that quickly becomes obsolete. This will help ensure currency of the data dictionary and will facilitate accessibility to all users by eliminating the need to generate and distribute paper documents.

As the single, comprehensive source describing the contents of the knowledge base, the data dictionary will be the basis for reference in tools that are being developed for accessing or browsing the knowledge base. This could be accomplished by implementing hypertext links from the browser tool into the data dictionary, so that when browsing the contents of the knowledge base, the user could immediately retrieve the associated description contained in the data dictionary. Similarly, the data dictionary could contain hypertext links to the browser, so that a user can read the data dictionary and immediately jump to viewing the dataset.

CONCLUSIONS AND RECOMMENDATIONS

The data dictionary provides a useful, accessible, and current catalog of the contents of the CTBT knowledge base. It reflects the ongoing refinements and enhancements to the knowledge base. The data dictionary defines basic datatypes that encompass all existing and anticipated data requirements; these datatypes provide the framework for documenting datasets. The data dictionary is in draft form and comments are welcome.

In order for the data dictionary to maintain its objectives for currency and usefulness, two things must happen. First, the authors have the responsibility of updating the data dictionary as new datasets are added and as refinements are made to the knowledge base. Secondly, and equally important, it is imperative that users provide feedback evaluating its usefulness.

REFERENCES

References that may be useful to developers or users of datasets.

Anderson, et al; *CSS Version 3 Database: Schema Reference Manual*, Science Applications International Corporation, September 1990. Describes the schema of the Version 3.0 database standard for data and software at the Center for Seismic Studies. When the phrase "CSS 3.0 format" is used, this is the document that is being implicitly referenced.

Armstrong, H., and Keyser, R.; *ARC/INFO Dataset Delivery Specifications*, memo dated December 10, 1996. Provides a description of what to include with delivery of research reference datasets that are in ARC/INFO coverage format, includes samples.

Hipp, J. et al; *Knowledge Base Interpolation of Path-Dependent Data Using Irregularly Spaced Natural Neighbors*, Sandia National Labs, 1996; published in *Proceedings of the 18th Annual Seismic Research Symposium on Monitoring a CTBT, 4-6 September 1996* (Editors: Lewkowicz, McPhetres, Reiter), 17 July 1996.

Shepherd, E. R., R. Keyser, H. Armstrong, E. Chael, C. Young (1997). Data Dictionary for CTBT Knowledge Base, Revision 1, DRAFT. Sandia National Labs, April 21, 1997.

References used in developing the contents of the knowledge base.

Devlin, B., *Data Warehouse from Architecture to Implementation*, Addison-Wesley, Reading, Massachusetts, 1997.

Federal Geographic Data Committee (FGDC), "Content Standards for Digital Geospatial Metadata," June 8, 1994. URL's:http://www.fgdc.gov/, http://www.fgdc.gov/Metadata/metahome.html.

HQ-AFTAC, "System Requirements for the NEMS (DRAFT)." 17 June 1996.

US-DOE, "Knowledge Base Product Definition, Phase 0, Revision 8." March 1997.

USNDC, "USNDC Knowledge Base Working List" revision: 21 Jan 97 (date of fax receipt, from Mark Woods [HQ AFTAC/TT] to Hillary Armstrong).