PROPOSED CONCEPTUAL REQUIREMENTS

for the

CTBT KNOWLEDGE BASE

Ralph G. Keyser, Senior Member of Technical Staff
and
Hillary M. Armstrong, Member of Technical Staff

Sandia National Laboratories

work completed under DOE ST485D

The authors greatfully acknowledge the input of a large number of people from the following organizations: AFTAC, ARPA, DOE, LANL, LLNL, PNL, SAIC, and S-Cubed

July 31, 1995

ABSTRACT

The United States Government (USG) has indicated its support for a Comprehensive Test Ban Treaty (CTBT) by 1996, however, the existing verification systems will not provide monitoring capabilities at the level demanded by such a CTBT. Work is underway at the Department of Energy (DOE) and other agencies to improve the capabilities of the existing verification systems.

One area of research is focused on acquiring regional knowledge of the Earth to allow detection and identification the low magnitude events required by a CTBT. The problems associated with organizing, storing, and making this knowledge available to automated processing routines and human analysts are significant, and solving these problems is an essential step in ensuring that research results are smoothly transferred into the operational environment. The proposed knowledge base is an approach to solving the problems of knowledge storage in a CTBT system.

In addition to providing regional knowledge to automated processing routines, the knowledge base will also address the ad-hoc methods now used for knowledge storage. This will make the overall data processing system easier to maintain and tune since knowledge will be stored in a well defined location and not duplicated in multiple ad-hoc storage schemes.

The Proposed Conceptual Requirements Document for the CTBT Knowledge Base is a high-level requirements document intended to provide an overview of the scope and function of the knowledge base. In addition, the conceptual requirements document also provides examples of the data types envisioned for storage in the knowledge base. This conceptual requirements document is in a development phase and will continue to evolve along with other documents which will be developed during the analysis and design phases of this project.

Keywords: automated data processing, knowledge base, geophysical model, parameterized data

INTRODUCTION

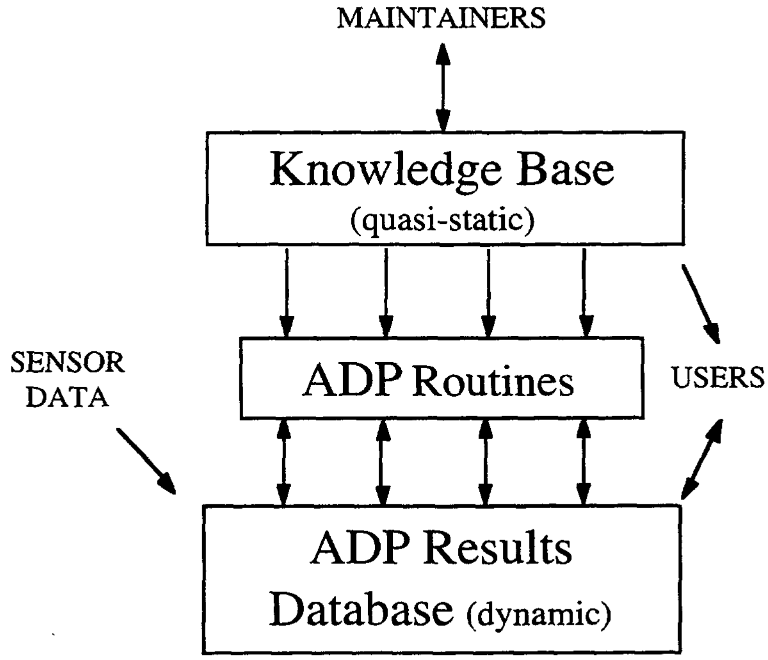

Fig. 1: The Knowledge Base

The proposed knowledge base will be a vital part of CTBT verification analysis. It will be a centralized storage area for the ADP parameters and geophysical data. As a central storage area, it will facilitate the integration of and access to information. The major verification tasks in which this information will be used are: phase detection and identification, association of detected arrivals with events, event location, event identification, and special event analysis.

The need for regional geophysical information is apparent from the desire to lower the global monitoring threshold for CTBT purposes. The automated data processing (ADP) routines need parameterized data such as travel time tables and attenuation tables in a form that is useful for routine data processing functions. The knowledge base will contain these regional parameters. The knowledge base will also need to manage underlying Earth model information for areas with no parameterized data, to allow refinement of existing parameterized data, and to support new algorithm development. Though both of these areas will be considered during analysis and design of the knowledge base, during implementation, first priority will be given to getting parameterized data into the knowledge base since that data is immediately useful to existing ADP routines.

The knowledge base will be a separate entity from the database containing the ADP processing results and the raw sensor data. An automated interface allowing the passage of parameters from the knowledge base to the ADP routines will be needed. The parametric data will also be available to the analysts and other users, so an interactive user interface will also be necessary (fig. 1). A data dictionary containing a preliminary list of these parameters can be found in Appendix 1.

The ADP results database changes almost constantly since it is continuously receiving new ADP processing results and sensor data. It is considered to be dynamic. By contrast, the knowledge base contents will change less frequently that of the ADP results database so it is considered to be quasi-static. Its parameters will be updated as the need arises. In order to make these updates, a maintenance interface will be needed (fig. 1).

This document discusses, at the conceptual level, the purpose of the knowledge base, who the customers are, where the knowledge base fits into the overall CTBT monitoring scheme, and the information and capabilities to be provided by the knowledge base. This document's primary purpose is one of initial content scoping in terms of data elements and functionality. This document is intentionally silent on the specific underlying hardware or software technologies that may be employed to support the knowledge base. (e.g., client server architecture, distributed database architecture, object oriented DBMS, relational DBMS, artificial intelligence, GIS, etc.) No inferences in the technology area are intended or implied. Furthermore, this document is not intended to be a functional requirements specification, a software requirements specification, a system design document, or an implementation plan. These will come later.

As discussed in the previous paragraph, this document provides the broad, high level scope of the knowledge base. Since its scope is necessarily broad, the knowledge base will need to have a phased implementation based on priority. The first priority will be to maximize performance in maintaining and providing the parameters to the ADP routines. Also high on the priority list is maintenance and provision of information currently used by the analysts and other users. Next on the list is maintenance and provision of parameters needed by ADP routines to be added in the near future. Following that is information which may be needed by the users in the future. Last on the implementation priority list is the storage of ancillary experimental data which may be of use in the future, but is currently not used in any algorithm.

Continued interaction with the customers (users) is vital to the success of any project. The Air Force Technical Applications Center (AFTAC) will be a primary user of the knowledge base. AFTAC is responsible for the development and operation of various US nuclear treaty monitoring systems including the prototype National Data Center (NDC), of which the knowledge base will be a part. The Advanced Research Projects Agency (ARPA) is another main knowledge base user. ARPA has a supporting responsibility to the Group of Scientific Experts (GSE). In that capacity, it is conducting the third Group of Scientific Experts Technical Trial (GSETT-3), which includes development of a prototype International Data Center (IDC), as well as detection and reporting capabilities to the Conference on Disarmament participants. The knowledge base will be an important source of regional information for the prototype IDC.

In addition to software development, the Department of Energy (DOE) Research and Development Laboratories are responsible for collection, integration, and synthesis of the information that will populate the knowledge base. DOE also has responsibility for research in support of CTBT monitoring efforts, so they will be users, as well as the developers, of the knowledge base. All of the expected CTBT technologies (i.e. seismic, hydroacoustic, radionuclide, infrasound, and on-site inspection) will be supported.

Another research and development user group is the community of organizations doing work on seismology and geosciences, such as the United States Geological Survey (USGS) and Incorporated Research Institutions for Seismology (IRIS). This community will both benefit from and contribute to information in the knowledge base. University and Industry research contractors comprise 2 large groups of users; the contractors under the joint agency Broad Area Announcement (BAA) managed by the Phillips Lab (PL), and the contractors working directly for AFTAC, ARPA, and the DOE Labs. Of particular note is Scientific Applications International Corporation (SAIC) which has a key programming / tool development role for both AFTAC (NDC) and ARPA (IDC).

ASSUMPTIONS

- The constant stream of information which the ADP routines receive in near real time will increase in size, due to the anticipated increases in the number and type of sensors being used.

- Lowering the minimum event threshold of interest, in order to meet U.S. Government detection goals, will increase the number of events to be examined.

- Regional variations in the geophysical character of the earth now need to be considered in the Earth models used in CTBT verification. Thus, the system will evolve from primarily teleseismic to primarily regional analysis.

- The knowledge base will need to cross traditional functional boundaries and be able to support all phases from detection through discrimination.

CONCEPTUAL REQUIREMENTS

o CONTENT SCOPING

1. **provide the parameters to process raw sensor data used in CTBT detection location, and discrimination tasks**
The knowledge base must effectively manage and provide the necessary parametric data to process data from the following sensor types:
seismic
radionuclides
hydroacoustic
infrasound

2. **support information at a wide variety of resolutions and extents**
The knowledge base will need to process and manage geophysical, and other information at a wide variety of resolutions and extents. Resolution is the minimum distance between 2 adjacent objects, or the minimum size of an object, which can be detected (or resolved). Extent is the amount of space or surface area which something occupies (i.e. the size of the area of interest) Resolution and extent usually, but not necessarily, have an inverse relationship; as extent increases, resolution typically decreases. This is very application-dependent and occurs for a variety of reasons including the fact that applications which look at very large areas typically do not need as much detail as a more specialized application which requires much greater detail (i.e., macro vs. micro). Since detail is usually costly to acquire and maintain, the ideal is to have adequate resolution and extent for the particular application. It is interesting to note that this discussion also holds for temporal resolutions and extents. The following are examples of geographic resolution and extent. A diagram showing how the sensors are connected together in an array at a particular seismic station would probably have a high resolution and relatively small extent. Whereas, a map showing the location of global seismic arrays would probably not need to have as high a resolution, and would definitely have a larger extent. Another example of the wide range of resolutions and extents can be found in geophysical models. A local geophysical model of the immediate vicinity of an individual event site would probably have a high resolution and small extent. A geophysical model encompassing all known sites of interest in a particular country might have to be at a lower resolution, due to data storage and manipulation limitations, and would have a larger extent.

3. **provide regional geophysical information in a practical, usable form**
To carry out the major verification tasks effectively, the automated processing routines will require information for regional distances. A major development will be to incorporate new information about the geophysical characterization for regions of monitoring interest. This development will be driven by the need for algorithm-specific parameters, such as regional travel times, regional phase decay rates, array beam sets, etc., based on the following two considerations. First, that such parameters be derived as closely as possible from

seismic recordings for source regions, receiver sites, and propagation paths of monitoring interest. Second, that to address the initial demands on the knowledge base and the development of improved parameters for source regions in general, a parallel effort will be the synthesis of earth models from which parameters of present and future algorithms can be derived.

o **FUNCTIONALITY**

1. **provide a smooth transition path to knowledge base use, so as not to impact present verification capabilities**
Installation of the knowledge base, particularly the ADP interface, must be smooth so as not to impact the present verification capability. The existing interfaces and data structures should initially remain intact; there need not be any changes to existing software. The initial knowledge base ADP interface should be backward-compatible, working with a snapshot of the parameter files. The snapshot would be made at a designated time and interval (e.g., every night at 11:00 P.M.) and would contain all parameter files necessary to run any of the ADP routines (fig. 2). This is the most static scenario. The next step would be a more closely-coupled interface where the ADP routines would work with a parameter extract. This extract would be made at the startup of a program and would contain the parametric data necessary to run that program. Finally, the most closely-coupled interface would allow interactive queries of the knowledge base as the program needed the data. (e.g., program extracts path-dependent travel times during execution.) The last two interfaces could be used together to provide the most flexibility. Eventually, a combination of the last two interfaces will be implemented as the primary ADP interface to the knowledge base (fig. 3).

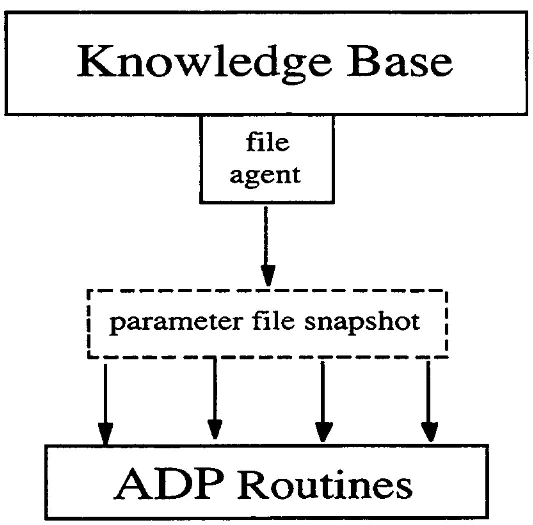

Fig. 2: Initial ADP Interface

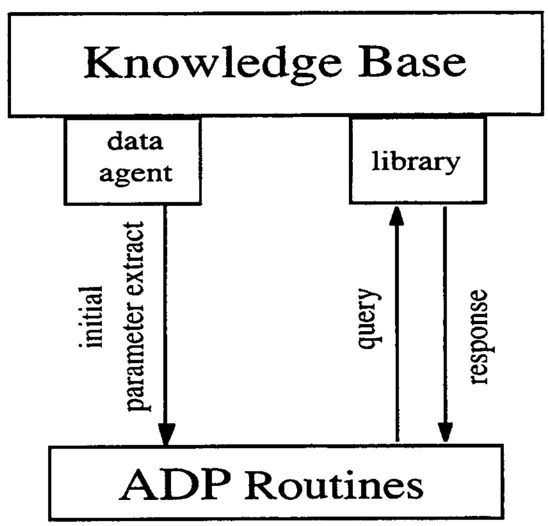

Fig. 3: Primary ADP Interface

2. **enhance the ADP environment through these automated interfaces**
The automated interfaces to the knowledge base (fig. 2 & 3). will help to improve the ADP environment by decreasing duplication of parameter data storage and standardizing parameter data access. Storing a parameter in one location only for a particular ADP routine will make the database easier to maintain and the changes easier to track. For example, it is critical to track what the value of a particular parameter was for a given run of

an ADP program. Having only one location of a parameter at a given time would make it easier to determine what the parameter value was at that time and, thus, what value was used for an ADP run at that time. The value of the parameter could be changed for subsequent runs, but would be time tagged and archived with other metadata to tie it to a particular run of the program. Development and eventual implementation of a standard way to access parameter data has the potential to improve program efficiency and make it easier to add new programs to the ADP environment. Given the assumption of an increase in the amounts and different types of raw sensor data to be handled by the ADP environment, this may prove to be a necessity.

3. **provide a user friendly, flexible, interface for accessing knowledge base contents**
The user interface to the knowledge base must be uncomplicated and easy to use, as it may need to operate in time critical situations. The analysts and other users need access to the same information as the ADP routines, as well as the models used to derive that information. The interface must also provide access to ancillary data necessary for special event analysis, that is, analysis of anomalous events which produced ambiguous results. The tools with which the users perform event analysis will have access to the necessary information in the knowledge base, however, development and maintenance of the tools themselves are explicitly out of scope of the knowledge base development project. The user interface to the database containing the ADP results and the raw sensor data is also out of scope.

o **MAINTENANCE AND DATA INTEGRITY**

1. **provide for creation, maintenance, and access to metadata**
Metadata is data about data. It provides the user with information about the origin, content, location, quality, condition, processing history and other characteristics of a given data element. The knowledge base needs to create, maintain, and make metadata easily accessible, as it is essential to preserving the integrity of the knowledge base. The following are some examples of how metadata would be used.

   o **determination of fitness for use**
   In order to determine which parametric data and models would be most appropriate to use for a particular event analysis, the analyst needs to know the accuracy (closeness to the truth) and precision (repeatability) at a given resolution and extent. For example, an algorithm may need to know the accuracy and precision of the various travel time computation options in order to decide which one to use.

   o **identification and description of raw sensor data source**
   Different Stations may have different equipment, sensor configuration, employee training level, data handling procedures, and communications protocol. Thus, the analyst will want to know which station the raw sensor data came from, the type and condition of the detection equipment, and other descriptive information about the station.

   o **identification and description of parametric data source**
   Metadata which identifies the source of models and parameters in the knowledge base are essential for maintaining a credible knowledge base with a verifiable scientific rationale. Metadata identifying the research report which generated a given parameter is one example of this.

- **knowledge base history**
  There may be a need to be able to examine the state of the knowledge base at some time(s) in the past to help explain analysis results and conclusions. For example, it may be necessary to determine what parameters values were used by a particular run of an ADP program in order to explain a result that is in conflict with results from other sources. Additionally, it may be necessary to trace the entire processing history of an event. One way to do this is by examining an audit trail pieced together from automatically collected metadata. There may also be parameters which vary with time (i.e., wind speed), so what the value was at a given time needs to be tracked. In order to be able to do these things, changes to the parametric and geophysical model data stored in the knowledge base need to be tracked as to when, why, and by whom they were made. If a parameter has more than one value at a given time in the data base (i.e. for different runs of an ADP routine) then information associating a given parameter value to a given program run must also be collected. Note that this last example applies to the knowledge base testbed only, as the operational knowledge base parameters will probably not be that dynamic.

- **provide feedback on ADP results**
  Metadata consisting of the error estimates for models and parameters in the knowledge base should be included for eventual use in error propagation calculations in the ADP algorithms. This will result in quantitative error analysis of results produced by the ADP algorithms that will aid the decision processes of later ADP algorithms, analysts, or other users.

2. **provide an extensible design**
   The knowledge base design needs to support development of new algorithms on the NDC testbed. It must be able to easily accept new data types, data structures, and models as they are developed. It must also accommodate ancillary experimental data which may be of use in the future, but is currently not used in any algorithm. Thus the representation, storage, and maintenance of geophysical knowledge in both parametric equation and 3-D Earth model forms will be necessary in the knowledge base.

3. **provide a structured maintenance environment for the enhancement, or evolution, of the knowledge base**
   Maintenance should follow established procedures in order to update and refine the knowledge base in an efficient, controlled, and organized way; configuration management is key here. A maintenance browser interface, with automated updating procedures and change recording capabilities may prove to be essential for this task. To assess the impact of updates and other maintenance tasks, system performance analyses will need to be regularly executed and archived (fig. 4). Note that although this is shown for the operational system only in fig. 4, there will be a similar feedback loop for the testbed system. The following are some of the maintenance tasks which may be required. In order to keep up with improving knowledge about the structure of the Earth, existing regional geophysical models will need to be replaced with new and improved models as they are developed. Additionally, newly created regional models will need to be integrated. In order to preserve knowledge base integrity, off-line Q.C. results will need to be incorporated according to some established procedures. Effective maintenance, together with metadata management, is essential to protect the integrity of the knowledge base and help to ensure that it remains a useful tool.

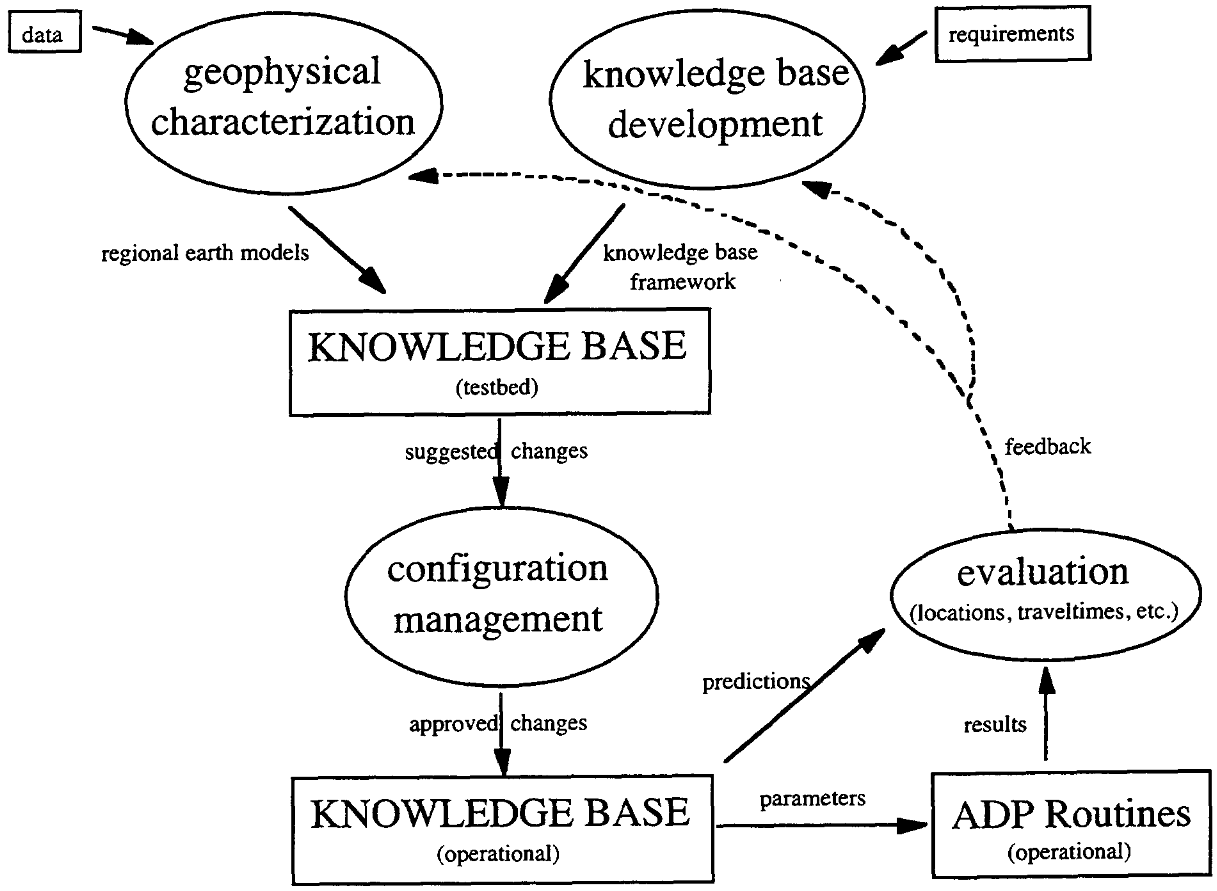

Fig. 4: Evolution of the Knowledge Base

Appendix 1

Knowledge Base Data Dictionary

July 31, 1995

At the heart of the CTBT Knowledge Base are the types of data to be included. This data dictionary is an attempt to list these data types and begin to group those data types into logical classes of data.

PATH-DEPENDENT INFORMATION

This class of information includes knowledge that depends on the path between a source event and the receiver. It will include information such as travel times and slowness along with correction factors for amplitudes, azimuths, etc. The knowledge base will support information at differing resolutions depending on the level of information available for the region. The knowledge base will also be able to support path specific information that varies by frequency or time.

Sample Data Types

1. Travel Time Table
2. Travel Time Corrections
3. Slowness Table
4. Slowness Corrections
5. Azimuth Corrections
6. Amplitude Corrections
7. Phase Blockage information
8. Discriminant Selection Information

ALGORITHMIC / PROGRAM-SPECIFIC PARAMETERS

This is information needed by a specific program or type of algorithm. It will support a wide variety of name-value pairs or vectors of name-value pairs that provide parameters for use by programs in storing information such as filter parameters, beam sets, signal detection parameters, measurement recipes, thresholds, etc.

Sample Data Types

1. Beam sets
2. Beam parameters

3. Signal Detection parameters
4. Channel sets
5. Filter sets
6. Filter parameters
7. Processing orders
8. Measurement recipes
9. Thresholds
10. Magnitude sets
11. Magnitude parameters
12. Noise measurement parameters

GEOPHYSICAL INFORMATION

This information class includes earth modeling information about a given point or region. It includes items such as tectonic region information, density and velocity data, knowledge about topography and geologic structure, and detailed data on seismicity in the region and its source.

Sample Data Types

1. Geographic Region
2. Tectonic Character
3. Density
4. Velocity
5. Q Information
6. Geologic Structure
7. Seismicity
8. Topography / Bathymetry
9. Depth to Moho
10. Depth to basement
11. Heat Flow
12. Gravity

13. Ocean Current

14. Reference Event Data including waveforms and event analysis

15. Sensor /Station Information

16. Wind

METADATA

This is "data about data". It provides the user with information about the origin, content, location, quality, condition, processing history and other characteristics of a given data element.

Sample Data Types

1. Quality

2. Origin

3. Processing history

GEOGRAPHIC INFORMATION

This information can be associated with a particular area of the earth. It may be useful as reference material.

Sample Information Types

1. Maps

2. Overhead Images

3. Mine Schedules

4. Demographics

5. Geopolitical

6. Cultural

7. Meteorological