ANNUAL REPORT ON TECHNICAL DOCUMENT INDEXING

(TDI)

Prepared for

Nuclear Regulatory Commission
Contract No. NRC-02-88-005

CNWRA MAJOR MILESTONE 20-3702-072-010

Prepared by

Sharon McFaddin

CENTER FOR NUCLEAR WASTE REGULATORY ANALYSES

SAN ANTONIO, TEXAS

September 9, 1991

1. INTRODUCTION

The following statistics, descriptions of current and planned modifications to TDI, and data to be added are provided to reflect the progress in the Center during FY91 in the subject area. Changes to the operating procedures in the form of a description of the TDI Search Facility are also attached (Attachment A).

Additionally, the U.S. NRC NUDOCS and other document reference databases have been accessed during the past year by the Center's staff. The staff continues to review documents and download specific sections of full text of HLW and related documents to support their research, technical assiatance, and, more specifically, Systematic Regulatory Analysis (SRA) tasks. The TDI system contains the headers, abstracts, and technical reviews for most of the document references used in these projects. Within the TDI bibliographic header index are stored all of the document references for the SRA data loaded using the Program Architecture Support System (PASS) in the Program Architecture Database (PADB).

It is becoming increasingly important for users of PASS who encounter TDI document reference to be able to have direct network access to NUDOCS to view/download the full text documents. At the present time, NUDOCS is only accessible via a dial-up telephone modem.

2. STATISTICS FOR DOCUMENTS, ABSTRACTS, AND REVIEWS PROCESSED

In the period October 1, 1990, to September 9, 1991, a total of 4400 document entries were added to the TDI database to bring the overall total to 10,207. Of this total, 837 documents were Regulatory Information Distribution System (RIDS) entries. The large majority of documents added were technically related to the National Institute of Science and Technology (NIST) and Oak Ridge National Laboratory (ORNL) databases (Section 4.) and totaled 3563 with an additional 1254 related abstracts. Approximately 170 reviews are contained in the database for the Geologic Setting, Engineered Barrier System, and Repository Design, Construction, and Operations Program Elements.

The Table I provides a monthly distribution and annual summary of the volume of TDI documents added to the database:

Table I. MONTHLY DISTRIBUTION AND ANNUAL SUMMARY OF TDI DOCUMENTS

MONTH/YEAR	TOTAL DOCUMENT QUANTITY	RIDS	TECHNICAL	ABSTRACTS
10/1990	184	83	101	0
11/1990	2762	0	2762	1087
12/1990	167	130	37	0
01/1991	81	64	17	0
02/1991	112	87	25	0
03/1991	186	82	104	0
04/1991	143	59	84	0
05/1991	270	90	180	0
06/1991	126	55	71	33
07/1991	221	107	114	81
08/1991	125	65	60	45
09/1991 (9 days)	23	15	8	8
TOTAL	4400	837	3563	1254

3. TDI MODIFICATIONS

During the last year, major improvements have been made to the TDI system in the search and retrieval of documents, and the methods of storing abstracts and reviewing data.

3.1. Retrieval of Documents

Enhancements were made to permit entry of compound search requests, including Boolean operators. This change in the retrieval logic permits the user to specify a very narrow or a very inclusive search request, and to modify and adapt the search request depending upon intermediate results. Thus, the user is able to begin searching with a very inclusive request, and then adapt that request to narrow it as the search progresses. This results in a greatly enhanced document search capability. A full description of this implemented capability is contained in Attachement A. It also covers the same capability which was implemented for the Center QA and Correspondence document indices. Documentation of the entire TDI system is in process and will be delivered during FY92.

3.2. Method of Storing Abstract and Review Data

The method of storing textual data such as abstracts and reviews has been changed to conform to PASS Version 2.0 SRA standards. This resulted in faster access and greater efficiency in the use of disk storage.

4. PLANNED MODIFICATIONS

There are no major TDI software system modifications planned for the next year.

4.1. NIST Database

The document index created by NIST has been converted to TDI input format and is loaded in the database. This provided access to an additional 1086 documents. The document review data for the NIST index has also been loaded and can be viewed. Additional keyword searches will be added with report formats defined for specific queries of this data.

4.2. Geochemical Literature References

Geochemical Literature References (GEOCHTA) is a database of geochemical literature references with review comments. It is a subset of the Waste Management Information System (WMIS) which resides at ORNL on an IBM 3033 using INQUIRE software from INFODATA. A copy of GEOCHTA was obtained on magnetic tape and installed in the TDI during the past year. Additional query and report capabilities will be added as necessary to accommodate searches of this data.

5. DATA BASES TO BE ADDED

The primary backlog of data to be added during FY92 are the results of ongoing literature reviews in several research projects and technical assistance tasks.

ATTACHMENT A

TDI SEARCH FACILITY

INTRODUCTION

There are three major subsystems in PASS/PADB which are intended to track documents. These document tracking subsystems are header index type systems in that they do not contain the full text of the documents but only enough identifying information to permit a user to identify and locate them. These three document tracking systems were implemented as distinct subsystems because of the differences in the information which must be maintained for the different types of records. Functionally, however, they are very similar in terms of their overall capabilities and user interface.

TDI	The Technical Document Index tracks technical documents which are in the CNWRA library or are referenced in bibliographies or literature reviews.
CSP	The Correspondence Index tracks CNWRA official correspondence documents.
QAR	The Quality Assurance Records subsystem tracks technical documents or correspondence which have been designated as a Quality Assurance Record.

These three records management subsystems are introduced separately because the user must understand the type of information contained in each subsystem and must select the appropriate one when searching for documents. All three records management systems are grouped as a single selection entry (PF-7) on the CNWRA MAIN MENU. Pressing PF-7 from this menu causes the CORRESPONDENCE/TDI/QA RECORDS Main Menu to be displayed. From this menu, the user may select search functions or other functions for any of the three records management subsystems.

GENERAL INFORMATION ABOUT THE SEARCH FACILITY

The CORRESPONDENCE/TDI/QA RECORDS Main Menu includes three selection items (PF-2, PF-5, and PF-8) which permit the user to search for documents in the TDI, Correspondence, and QA Records subsystems, respectively. By pressing the appropriate PF-key, the search facility for the desired records management subsystem may be selected. The search facilities for the three records management subsystems are quite similar, differing only in the data fields which may be searched.

Records may be found in the Correspondence Index subsystem by searching for specific data in the following data fields:

DOCument number;
SUBJECT Code;
SUBject;
PROJect number;
DATE of the document;
AUthor;

ADDRessee.

Documents in the QA Records subsystem may be found by searching for specific data in the following fields:

DOCument number;
DATE of the document;
SUBJECT Code;
AUthor;
SUBject;
TITle;
PROJect number.

Documents in the TDI subsystem may be found by searching for specific data in the following fields:

DOCument number;
SUBJECT Code;
REPort number;
DATE of the document;
AUthor;
AFFiliation of the author;
EDitor;
TRanslator;
COmpiler;
SUBject;
TITle;
SOurce;

When the search facility is selected for the desired subsystem, a screen will be displayed which permits entry of the search parameters. Each search command that is entered is called a query predicate. A search may be composed of a single query predicate or multiple query predicates which are connected by the "AND or "OR" logical operators. In general, the user must supply three pieces of information for each query predicate:

SEARCH COMMAND - This command tells the search facility what the user wants to do. This field may contain one of three values:

1) The word "SEARCH" to begin a new search. The "SEARCH" command is only valid for the first search predicate in a query.

2) The word "AND" to connect the current search predicate in an "AND" relationship with the results of all prior search predicates.

3) The word "OR" to connect the current search predicate in an "OR" relationship with the results of all prior search predicates.

SEARCH TERM - This is the actual data which the user is trying to locate in the records management database. There are certain rules which apply to the search data depending upon which data field is expected to contain the information. For example, when searching for a name in the author, editor, translator, compiler, and similar fields of records in the TDI subsystem, the name data must be entered in upper and lower case with the last name appearing first. Similarly, when searching for a date, the date must be entered in the format, "yyyymmdd".

FIELD NAME - This is the name of the data field which will be searched to find the data. If the field name is omitted, then the search facility will default to looking for the search term in all fields which have been automatically processed for keywords. (See below for SPECIAL CONSIDERATIONS FOR KEYWORD SEARCHES)

ENTERING MULTIPLE SEARCH PREDICATES

The search facility permits the user to enter multiple search predicates. As each predicate is entered, the system will determine the number of records which satisfy the current search criteria and the number which satisfy the combination of all search predicates which have been entered for the current query. The number of records which meet the cumulative criteria of all search predicates is indicated in a line in the middle of the screen. For example, the following message would indicate that a cumulative total of 27 records have been selected by the combination of all of the query predicates in the current query:

27 RECORD(s) SELECTED BY THIS QUERY

As each query predicate is processed, the query parameters are displayed in a scrollable table of query predicates along with an indication of the number of records which satisfied that particular query predicate. For example, the following table illustrates a four-predicate query which resulted in a single document being selected:

_	1 SEARCH	(Chung* IN Author)	35 HITS
_	2 AND	(earthquakes IN title)	13 HITS
_	3 AND	(effects IN title)	95 HITS
_	4 AND	(Carpenter* IN Author)	3 HITS

After a query has been formulated, individual predicates may be selected and modified to change and re-execute the query. A selection field is provided at the left of each query predicate in the query table and the following values may be entered in the selection field for any of the prior query predicates:

Entering an "A" in the selection field for a particular query predicate causes the current query predicate to be inserted AFTER the selected query predicate.

Entering a "B" in the selection field for a particular query predicate causes the current query predicate to be inserted BEFORE the selected query predicate.

Entering a "C" in the selection field for a particular query predicate moves the selected predicate into the top part of the screen where it may be modified and executed again.

Entering a "D" in the selection field for a particular query predicate deletes the selected predicate.

SEARCH TERMS

The search term is used to define the actual data which the user wishes to find.

SIMPLE SEARCH TERMS

Most search predicates will use simple search terms composed of single words. For example, one could search for the word "environmental" by entering that word in the search term and then entering an appropriate field name to define the field in which to search for "environmental".

SEARCH TERMS WHICH USE WILD CARD CHARACTERS

Search terms may also contain an asterisk (*) which serves as a "wild card" character. For example, the search term "Chung, D. H." would look only for names which exactly matched the search term in the specified field. But using the wild card character permits searching for inexact matches. For example, the search term "Chung*" would look for all names beginning with the characters "Chung" regardless of any following characters or initials.

COMPLEX SEARCH TERMS

Slightly more complex search terms may be constructed by joining two words with a logical operator such as "AND" or "OR". For example, one could enter the search term "ENVIRONMENTAL AND CONTAINMENT" and an appropriate field name to find all records which contained both the words "ENVIRONMENTAL" and "CONTAINMENT" in the specified field. Similarly, if one entered the search term "ENVIRONMENTAL OR CONTAINMENT" and an appropriate field name, the query processor would retrieve all records which contained either the words "ENVIRONMENTAL" or the word "CONTAINMENT" or both words in the specified field.

SEARCH TERMS FOR NAMES

Search terms for names are compared to the data base records on a character for character basis. The search facility assumes that an exact match is desired. Upper and lower case are significant when searching for names, as is punctuation and the number of spaces between words. Therefore, the search terms for names must be entered exactly as they would appear in the database. By convention, this means that the following rules should be observed for name search terms:

1) The first letter of each word should be capitalized.

2) The last name should appear first.

3) The last name should be followed by a single comma.

4) The first name and initials should follow the last name and comma.

5) Initials, if any, should be followed by a single period.

6) A single space should follow each period (.), each comma (,) and each word which does not end in punctuation.

If the first name and initials are not known, then the name may usually be found by entering the last name, followed by a wild card character (*). This type of generic search for a name, however, may retrieve more entries than actually desired. For example, the search term "Smith*" would retrieve all last names which begin with the character string "Smith". The answer set for this search term would include all records for names in which the last name was "Smith", "Smiths", "Smithers", "Smithson", etc.

SEARCH TERMS FOR DATES

Some special search facilities have been implemented for dates. In searching for a date, it is important to remember that all dates are stored internally in the database in a sortable form. That is to say that the internal format of all dates is "yyyymmdd" where:

yyyy is a 4 digit year
mm is a 2 digit month
dd is a 2 digit day

In the internal representation of a date, the year subfield should always be present, but the month and/or day may be omitted by entering a "00" for the month and/or day.

Thus, to search for the date May 23, 1991, one would enter the search term as "19910523".

SIMPLE DATE SEARCH TERMS

Many date queries may be accomplished using simple date search terms composed of a single date. For example, one could enter "19910523" in the search term and "DATE" in the field name to search for all documents with a publication date of May 23, 1991. Similarly, one could enter "19910500" in the search term and "DATE" in the field name to search for all documents with a publication date of May, 1991 (i.e. the day subfield was not specified when the publication date was entered).

DATE SEARCH TERMS WHICH USE WILD CARD CHARACTERS

Date search terms may also contain an asterisk (*) which serves as a wild card character. For example, if one entered "199005*" in the search term and "DATE" in the field name, all documents with a date of May 1990 regardless of the contents of the day subfield would be retrieved. Similarly, if one entered "1990*" in the search term and "DATE" in the field name, all documents with a date of 1990 regardless of the contents of the month or day subfields would be retrieved.

Wild card search terms should be used very carefully with date fields to avoid retrieving very large numbers of documents.

COMPLEX DATE SEARCH TERMS

Three additional features are provided for date search terms which permit retrieval of records BEFORE a specified date, AFTER a specified date or BETWEEN two dates.

SEARCH TERMS BEFORE A SPECIFIED DATE

One may enter the keyword "BEFORE" followed by a date, with or without a wild card character, to retrieve all documents with dates which preceded the specified date. For example, if one entered "BEFORE 19900523" in the search term and "DATE" in the field name, all documents published prior to May 23, 1990, would be retrieved.

SEARCH TERMS AFTER A SPECIFIED DATE

One may enter the keyword "AFTER" followed by a date, with or without a wild card character, to retrieve all documents with dates which follow the specified date. For example, if one entered "AFTER 19900523" in the search term and "DATE" in the field name, all documents published after May 23, 1990, would be retrieved.

SEARCH TERMS BETWEEN TWO SPECIFIED DATES

One may enter the keyword "TO" between two dates (either of which may or may not contain a wild card (*) character) to retrieve all documents with dates which fall between the first and second date. For example, entering "19900523 TO 19900601" in the search term and "DATE" in the field name, would retrieve all documents which contained dates between May 23, 1990 and June 1, 1990.

FIELD NAMES

The following table indicates the various acceptable field identifiers which may be entered in the search predicate for each of the three subsystems in the PASS/PADB search facility. The capitalized portion of the field identifier is what the program is actually checking, so abbreviations to that level are permitted. For example, specifying "DOCUMENT" is the same as specifying "DOC". Similarly, specifying "AU" is the same as specifying "AUTHOR".

Records may be found in the Correspondence Index subsystem by searching for specific data in the following data fields:

DOCument number;
SUBJECT Code;
SUBject;
PROJect code;
DATE of the document;
AUthor;
ADDRessee.

Documents in the QA Records subsystem may be found by searching for specific data in the following fields:

DATE of the document;
AUthor;
SUBJECT Code;
SUBject;
TITle;
PROJect number.

Documents in the TDI subsystem may be found by searching for specific data in the following fields:

DOCument number;
SUBJECT Code;
REPort number;
DATE of the document;
AUthor;
AFFiliation of the author;
EDitor;
TRanslator;
COmpiler;
SUBject;
TITle;
SOurce.

If no field name is entered, the system will assume that a general keyword search is desired, and it will retrieve all occurrences of that keyword in all fields which have been automatically processed for keywords. This type of general keyword search should be used with caution for two reasons. First, only selected fields are processed for keywords. For example, none of the name fields in the TDI or QA Records subsystems are processed for keywords. Therefore, searching for a name without specifying the field to be searched will not find any documents in these subsystems. Second, searching for a common word without specifying which field to search will result in all occurrences of that keyword being retrieved. This may result in a very large number of documents being retrieved, which can seriously degrade the performance of the system.

SPECIAL CONSIDERATIONS FOR KEYWORD SEARCHES

Certain data fields are automatically processed to extract keywords at the time that the records are entered into the database. This automatic keyword processing involves scanning the data field and extracting all "significant" words which are then stored as keywords. Common prepositions and conjunctions such as "and", "or", "of", "before", "to", etc. are ignored in this processing. Keywords are automatically raised to upper case and therefore are not case sensitive. Searching for "CONTAINMENT" or "Containment" or "containment" are all equivalent. Thus, "significant" words in selected data fields may be selected through this keyword method.

In the TDI subsystem, the following fields are automatically processed for keywords:

TITLE;
SUBJECT;
SOURCE;
AFFILIATION.

In the QA Records subsystem, the following fields are automatically processed for keywords:

TITLE;
SUBJECT.

In the Correspondence subsystem, the following fields are automatically processed for keywords:

AUTHOR;
ADDRESSEE;
SUBJECT.

When retrieving a search term which has been processed as a keyword, the search term may be entered in either upper or lower case and the appropriate field name should be entered. If no field name is entered, the system will search for all occurrences of that keyword in all keyword fields. This may significantly increase the size of the answer set as well as the time required to retrieve and process it.

In general, wild card characters (*) should be avoided when searching fields which have been processed for keywords. This is because there is a very large number of keyword entries in the database, and the retrieval and processing time may be excessive if the wild card (*) character is used in the search term.