ANNUAL REPORT ON TECHNICAL DOCUMENT INDEXING

(TDI)

Prepared for

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

Prepared by

Robert L. Marshall

Center for Nuclear Waste Regulatory Analyses
San Antonio, Texas

August 1993

CONTENTS

TABLES

1 INTRODUCTION

The Technical Document Indexing (TDI) system contains the headers, abstracts and technical reviews for most of the document references used by the CNWRA staff. The Program Architecture Support System (PASS) utilizes the TDI bibliographic header to store all document references for the Systematic Regulatory Analysis (SRA) data contained in the Program Architecture Database (PADB). The staff will continue to review documents and download specific sections of full text from high-level waste (HLW) program and related documents in support of their research, technical assistance and SRA tasks. Other relevant document databases have been loaded into the TDI to augment its data and facilitate its utility.

It is important for PASS users, who encounter TDI document references, to access the U.S. Nuclear Regulatory Commission's (NRC) NUDOCS, in order to view/download the full text documents. Although the CNWRA staff has accessed NUDOCS and related document databases during the past year, they are currently accessible only via a dial-up telephone modem.

Section 2 contains statistics relative to the nature and volume of TDI documents that have been added to this database. Section 3 offers comments concerning the current status of the TDI, and it provides some commentary on anticipated changes for the following fiscal year. In addition, a description of the TDI Search Facility (which can serve as a brief user's guide) is included as Appendix A of this report. The CNWRA has implemented two other document indexing systems almost identical to TDI for control of correspondence and QA records which are referenced in Appendix A.

2 STATISTICS FOR DOCUMENTS, ABSTRACTS, AND REVIEWS PROCESSED

In the period from August 18, 1992 to August 11, 1993, a total of 2,224 document entries were added to the TDI database. This has resulted in an 18 percent increase in the number of entries this fiscal year–from 12,318 to 14,542 entries. Of these, 1,003 documents were provided by the Regulatory Information Distribution System (RIDS). The majority (i.e., 1,221) of the documents were technical documents identified by CNWRA investigators as being pertinent to their technical work. Those documents were acquired by CNWRA and indexed in the TDI prior to their use. Of these documents, 282 had abstracts in the original document. Approximately 171 reviews are contained in the database that were performed by the Geologic Setting, Engineered Barrier System and Repository Design, Construction and Operations Program Elements.

Table 1 provides a monthly distribution and annual summary of the TDI document volume added to this database.

Table 1. Monthly distribution and annual summary of TDI documents

MONTH/YEAR	TOTAL DOCUMENT QUANTITY	RIDS	TECHNICAL	ABSTRACTS
08/1992*	35	8	27	6
09/1992	301	118	183	60
10/1992	143	94	49	12
11/1992	115	53	62	5
12/1992	157	89	68	20
01/1993	194	69	125	23
02/1993	183	78	105	17
03/1993	285	129	156	29
04/1993	176	122	54	15
05/1993	197	84	113	40
06/1993	179	70	109	27
07/1993	158	58	100	8
08/1993**	101	31	70	20
TOTAL	2224	1003	1221	282

*8/18/92-8/31/92
**8/1/93-8/11/93

3 TDI STATUS

During the last year the TDI system was used primarily in an operational mode, as intended, without any significant modifications. A reporting facility capable of listing the previous month's acquisitions was added. Operational problems related to database performance (e.g., poor search times) were addressed and corrected in a timely and efficient manner as part of normal database maintenance.

The CNWRA is evaluating its overall document control requirements as part of the planning process for internal computer operations and acquisitions. Also, as part of Task 6 of the Center Operations Element work, the CNWRA has completed the required definition for the Division of High Level Waste Management (DHLWM) Technical Reference Document Database System (TDOCS) (Johnson, 1993) and will soon begin design work on the TDOCs system. Alternatives identified by this design effort will be evaluated during the course of the next fiscal year, and their implementation will be contingent on the results of the evaluations.

Until the major alternatives are implemented, it is not expected that all of the CNWRA literature reviews and bibliographic information will be loaded in the TDI in a timely manner. At this time, literature reviews from a number of projects are maintained on stand-alone PC databases. While these individual bibliographic files support the needs of the individual and the specific project, they are, for the most part, not available for other staff to review online in TDI. Although progress has been made this fiscal year in these areas, more is expected in FY94.

4 REFERENCES

Johnson, R.D., J.H. Cooper, C. Moehle, E. Harloe. 1993. *Technical Reference Document Database System (TDOCS) Requirements Definition*. Center for Nuclear Waste Regulatory Analyses: San Antonio, TX.

APPENDIX 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 TDI/CORRESPONDENCE/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 TDI/CORRESPONDENCE/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.

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;

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.

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

A reset key (PF5) is provided to clear the current query and begin a new one. 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".

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.

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:

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

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

AUTHOR;
ADDRESSEE;
SUBJECT.

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

TITLE;
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.