CENDI

CENDI AGENCY INDEXING SYSTEM DESCRIPTIONS:
A BASELINE REPORT

Sponsored by the
CENDI Subject Analysis and Retrieval
Working Group

Prepared by
Gail Hodge
Information International Associates, Inc.
Oak Ridge, Tennessee

April 1998

CENDI Subject Analysis and Retrieval Working Group Members

Defense Technical Information Center (DTIC)
Tanny Franco
John Dickert
Annie Washington
Department of Energy, Office of Scientific and Technical Information (DOE/OSTI)
Barbara Bauldock
Nancy Hardin
National Aeronautics and Space Administration (NASA)
Michael Genuardi
National Air Intelligence Center (NAIC)
Mindy Sorenson
Richard Hammond
National Library of Medicine (NLM)
James Marcetich
National Technical Information Service (NTIS)
Kris Vajs (Chair)
Mona Smith
US Geological Survey/ Biological Resources Division (USGS/BRD)
Judy Buys

CENDI Secretariat
Gail Hodge

CENDI is an interagency cooperative organization composed of the scientific and technical information (STI) managers from the Departments of Commerce, Energy, Education, Defense, Health and Human Services, Interior, and the National Aeronautics and Space Administration (NASA).

CENDI's mission is to help improve the productivity of Federal science- and technology-based programs through the development and management of effective scientific and technical information support systems. In fulfilling its mission, CENDI member agencies play an important role in helping to strengthen U.S. competitiveness and address science- and technology-based national priorities.

TABLE OF CONTENTS

EXECUTIVE SUMMARY

1.0   BACKGROUND

In August, 1997 the CENDI Principals and Alternates approved the formation of a Subject Analysis and Retrieval (SAR) Working Group. Its charter is to provide input to CENDI on opportunities for cooperation and education in the areas of indexing, thesaurus management, and retrieval. As a means of introduction and to give a baseline for the SAR WG's activities, each agency was asked to provide a brief description of the indexing performed by the agency and the major concerns related to indexing/subject access.

Over the course of a face-to-face meeting and several follow-up teleconference calls, the agencies described their indexing systems, their staffing, training, future plans, and issues. The commonalities are outlined in the "Summary" section.

2.0   AGENCY DESCRIPTIONS

2.1   National Technical Information Service (NTIS)

NTIS takes information from many of the CENDI agencies and other federal agencies, and creates a large database with the purpose of providing access to government information by the public. (Approximately 21,000 of the 62,000 documents indexed this year will be from CENDI agencies.) NTIS melds the records from other agencies into their own system. To provide subject access to this broad range of material, NTIS uses seven to eight relevant thesauri in specific subject areas, including those from the CENDI agencies and others, to adequately describe the material, and to ensure that the relationships between material from diverse agencies are manifested.

There was a recent attempt to create a single thesaurus by merging the electronic versions of the thesauri currently in use. However, when the project leader retired, the effort was not continued.

All indexing is performed internally. Indexing and cataloging are done separately. As acquisitions increase, particularly in the non-STI areas, the internal staff resources are being strained. NTIS is interested in automated tools that others are using.

The problem of indexing business literature in a scientific/technical environment have been highlighted by the International Tradebook Store project with the Department of Commerce. This project includes primarily policy documents. There was a need for additional vocabulary terms that were not contained in the previously used thesauri. NTIS has begun to use the Congressional Research Service thesaurus for business and policy related subjects.

2.2   Department of Energy, Office of Scientific and Technical Information (DOE OSTI)

OSTI has been the manager of STI for the whole of DOE since 1947. In addition to the DOE responsibilities, OSTI serves as the US representative to the International Nuclear Information System and the Energy Technology Data Exchange (ETDE). OSTI also serves as the ETDE

operating agent which means it collects input from the other countries, processes it, and returns the combined database to the member countries. OSTI is currently processing about 160,000 records/year.

OSTI is moving quickly to an open system Internet environment and away from paper. They have processed approximately 15,000-20,000 full text reports in electronic form back to January 1996.

Some cataloging/indexing is done by a local contractor. They also receive bibliographic records for journal articles electronically from the American Institute of Physics (AIP) and technical report citations from other government agencies.

The ETDE Thesaurus is used to index the documents. It was derived from the US DOE Thesaurus, but has been modified over the years. Recently, the ETDE and INIS thesauri have been made compatible, though they remain two separate thesauri with INIS a subset of the ETDE.

2.3   US Geological Survey/Biological Resources Division (USGS/BRD)

The USGS/BRD does not have a central STI database or publications process. They are developing a new system which is based on distributed resources and networking. The USGS/BRD has leadership responsibility within the federal government for the National Biological Information Infrastructure (NBII). Both the vocabulary and publications projects being developed within the new system will be components of the NBII.

The vocabulary will be used in several ways. It will be used to describe the USGS/BRD publications and other electronic resources such as data sets. Most of these resources will be cataloged and indexed by the resource creators/owners as part of a metadata initiative. Some older documents are being cataloged by a library technician. The vocabulary will also be used to describe the resources from other organizations collected under the NBII. The vocabulary will be provided as a search aid for the NBII and other biological, ecological, and environmental collections.

The vocabulary for biodiversity will be a joint project between the USGS/BRD and other organizations, both governmental and commercial. BRD is currently working with the California Environmental Resources Evaluation System (CERES). When the existing projects and thesauri were evaluated it was determined that CERES had already started a metadata subject description initiative that would fit into the work to be done at the federal level.

The CERES vocabulary is currently organized into nine hierarchies—natural resources, natural environment, demographics and infrastructure, boundaries, cultural resources, technologies, science (the disciplines), laws and regulations, and the human environment. Approximately 1,500-2,000 words have been organized under these hierarchies, many of them gathered by mining the Library of Congress Subject Headings, but then decomposing the pre-coordinated terms.

Another key approach to the NBII Vocabulary is that it will be very shallow, perhaps only three levels deep. Rather than develop an extensive hierarchy which will be difficult to maintain, since

no resources have been provided for maintenance, the vocabulary will be shallow with state and other local vocabulary clustered underneath. Also, at various points in the hierarchy, links will be made to more detailed thesauri. These thesauri may be from other government agencies, from within the USGS itself, from commercial organizations, or from non-profit groups. A major effort will be to determine how these links should be made. Discussions are planned with other groups interested in linked thesauri and standards for creating and using such a tool on the Web. A pilot project is planned.

Work continues to ensure that the high levels of the vocabulary are the appropriate ones for the NBII. It may be that Cultural Resources, which are not as important at the national level as they are at the local level, may not be a high level hierarchy. Biological modeling tools and techniques are a central component of the resources to be provided by the NBII. They may become a high level node since these need to be emphasized.

2.4   National Aeronautics and Space Administration, STI Program (NASA)

The NASA Center for AeroSpace Information (CASI) is part of the NASA Scientific and Technical Information (STI) Program. The STI databases contain over three million records. At least two million of them are technical reports and journal articles.

Documents are indexed using a controlled vocabulary—the NASA Thesaurus. (While there is a field available for uncontrolled identifiers, this field is not currently used for original indexing.) The NASA Thesaurus has been in existence since 1976 in its present form which includes term hierarchies, related terms, and cross references. It now contains approximately 17,700 terms, 4,000 USE references, and over 3,800 definitions . The scope of the Thesaurus and the STI databases are actually very broad covering the aerospace sciences, natural space science, and all supporting areas of physics, materials science, engineering, biology, etc.

In Volume 1 of the Thesaurus, terms are arranged in alphabetical order.   Similar to LeSH, the display format presents hierarchical information (referred to as 'generic structure'): for each term all levels of broader and narrower terms directly associated with that particular term are presented. Internal relationships between broader or narrower terms are indicated with indentations. Related terms, USE FOR references, and scope notes are also displayed.

NASA is currently preparing the 1998 edition of the Thesaurus which will incorporate definitions into the format of the main volume along with the proper upper/lowercase for the terms (previously presented in all uppercase). Volume 2-- the Access Vocabulary -- gives an alphabetical listing of terms (and USE reference) in each permuted form. This will be replaced with a KWIC-type index (rotated term display) for the 1998 edition. (The lexicographer is currently moving many of the permuted forms into the main thesaurus volume as USE references.)

The Thesaurus Edit System, used to maintain the NASA Thesaurus, was developed internally at CASI. It allows the lexicographer to call up any term and display the complete hierarchy, including all internal relationships. The lexicographer can directly add or delete broader, narrower, or related terms, definitions, scope notes, etc. Creation and modification dates, and other 'management' fields are automatically added to the term record. The system's dual screen

shows multiple hierarchies or parts of hierarchies so they can be reviewed against each other. The system was actually designed to handle more than one thesaurus, but NASA does not currently use it in this way.

The CASI indexers use the full document for abstracting and indexing with exceptions for cases where the original document is not available.

The majority of the documents are handled by the indexing staff with Machine Aided Indexing (MAI) support. The MAI process is an integral function of the Input Processing System (IPS) used at CASI. When the MAI button is clicked on the abstract screen, the text of the title and abstract are processed with an average three to six second response time. The result is a list of candidate terms from which the indexer can select. The MAI process also has a built-in spell check function which identifies words in the title and abstract that are not found in the MAI Knowledge Base.

The indexer can also select terms directly from an integrated online thesaurus. The indexing screen provides full display, search, and navigation capabilities for the thesaurus. The indexer can begin with the alphabetical list of terms and then move to the hierarchical display. The thesaurus can be accessed by browsing, by keying a truncated string, or by highlighting a string of characters in the abstract and then using the string to search the thesaurus. The indexer can then select and move terms to the working space for indexing. The thesaurus display, the title-and-abstract or MAI output can be displayed side-by-side with the indexing field so that the indexer has full flexibility to choose what is viewed on the screen at any given time. For documents that are received from the CENDI partners or purchased from a publisher or database producer, the indexer can also display the non-NASA index terms applied by the other organization.

Once an index term is selected, the indexer adds the term to one of two indexing fields-- Major Terms or Minor Terms. Major terms are indicative of the main theme of the document. Minor terms are of secondary importance.

NASA CASI has used Machine Aided Indexing (MAI) for many years. The MAI system does not perform full natural language processing or grammatical parsing. Instead it uses certain computational linguistic rules which process the text and give results approximating the results of full grammatical parsing-- but without the computational overhead. The rules have been developed to get at those features of text that have the most potential for representing indexable concepts.

MAI uses a large knowledge base (KB) of over 170,000 words and phrases. Maintenance is ongoing. The KB also contains other types of entries. Some entries in the KB function in coordination with the computational rules during the text analysis process to direct the creation of extended phrases; the KB plays a role in the parsing of the text-- it is not just a list of phrases with a straight look up. (There have been several articles published on the NASA MAI and how it works.)

In order to build the knowledge base, NASA CASI developed a statistically-based text analysis program (the KBB) which runs against large subpopulations of records in the STI database. Each subpopulation relates to a particular thesaurus concept. Computational parsing of the text occurs

and the system presents a ranked list of phrases that contain synonyms and variant expressions that can be reviewed by a subject analyst. Appropriate phrases are then moved to the knowledge base. NASA began using this statistically-based text analysis program in 1988.

Some of the open literature added to the database is not manually indexed. In the case of Compendex records, a Subject Switching routine is used to match terms, as best as possible, from the Engineering Information Thesaurus to those of the NASA Thesaurus. This output is supplemented with a MAI-type process for identifying additional terms. As part of the Compendex automatic indexing process, records where the output is identified as 'insufficient' are tagged so a manual review can be carried out.

NASA is improving some of the intelligence features that support the cataloging function in their input processing system. The old mainframe system had many rules that ran behind the scenes for quality assurance and data validation. The new system has some of them incorporated, but not all. In the area of abstracting and indexing, there are several improvements that could be made to the accessibility of the thesaurus.

About 35,000 technical reports and approximately 40,000 open literature items are added to the database per year. Approximately six full-time indexers provide the indexing and abstract writing and editing, and another two catalogers have been cross-trained to do indexing part time.

Currently, the main source for open literature is Compendex. As mentioned earlier, a combination of term switching and MAI is used to support the indexing of this material. Starting soon, NASA will be bringing in records from other sources such as the American Institute for Aeronautics and Astronautics (AIAA) with NASA indexing already applied. Compendex will then be used to 'fill-in' with records from peripheral subject areas.

Possible future plans for the Knowledge Base and the Thesaurus include the development of a more conceptual-network-type system that can be used within the retrieval interface.

There has been heavy use of the NASA Thesaurus via the Internet since an experimental HTML resource was created about four years ago. Unfortunately, this prototype electronic version has not been updated since its creation. There are many features that users would like to see in electronic form; with new features in place the electronic version would be an attractive alternative to the printed publication. Users want to display the full hierarchies, and provide an electronic equivalent to the Access Vocabulary, including embedded searching. The real test will be the Internet response time when going from view to view with these enhancements. In the interim, NASA has provided a PDF version of the Thesaurus on the Web in publication format. Despite the initial interest and use of the prototype online version, others still want the hard copy.

2.5   National Library of Medicine/National Institutes of Health (NLM)

NLM indexers use the Medical Subject Headings (MeSH) which currently contains about 18,000 unique subject terms. The indexers previously used a printed annotated alphabetical list. Increasingly, the indexers consult the vocabulary online. The list includes annotations. The numbers beside the terms are the MeSH tree numbers. MeSH is arranged into hierarchies or trees.

The indexers apply about eight to ten terms to each article depending on its length and how much coordination of terms is needed. MeSH is used first to match a text word in the title or abstract to the terms in the permuted vocabulary. Indexers may also look at previously indexed terms in MEDLINE.

The NLM online system is capable of multitasking. The record and the indexing system or MEDLINE can be on the screen at the same time. The MeSH is searchable as an inverted file. A fragment or text word search can be performed using the Elhill search command language (developed for MEDLINE). The MeSH file is also available in Folio Views in a hypertext-like file.

The indexer can interact with the MeSH file from within the indexing application. The indexer can neighbor terms and access the display or tree terms within the indexing application. Indexers can also display the MeSH annotations and other relevant information. In addition to the 18,000 terms in the MeSH vocabulary, there are also 100,000 supplemental chemical terms. These terms are indexed differently because the file contains a large amount of chemistry, in the areas of drug administration and diseases studied at the molecular and biochemical level. If the chemical term is not found, an entry is created by the indexer in the Supplementary Chemical list, rather than adding it into the MeSH thesaurus itself.

The Supplemental Chemical list may have broader chemical terms associated with them. These broad terms link into the thesaurus by mapping to thesaurus terms. The programming that builds the file actually adds to the thesaurus the terms to which the chemical term is mapped.

The MeSH headings appear after the MH tag. Some terms have a "/" and words in all caps. These are the subheadings or the qualifiers. Subheadings or qualifiers help to more fully describe the article. Qualifiers may include terms such as "diagnosis", "drug therapy," etc. Only certain qualifiers can be used with a given MeSH heading.

Major concepts are indicated by an asterisk. The major terms are included in the print Index Medicus. Those terms that do not have an asterisk are only available online. The Publication Type (PT) terms are also supplied by the indexer. At least one PT is required. PTs include journal article, letter, editorial, news item, etc. A class of articles called randomized controlled trials are also identified and entered.

Over 500,000 articles were indexed last year because of the data entry crisis the previous year. Typically, about 400,000 items are indexed from about 4,000 biomedical journals world-wide. The in-house staff of about 30 people do about 12 percent of the production. Eighty percent is done under contract. Eight percent is produced by foreign centers such as the British Library.

The in-house staff do training and reviewing of the contractors work. Each new indexer and each contractor is assigned a senior indexer in-house. As the indexer becomes more experienced there is less review; however, output is always spot-checked.

The training period for indexers is a formal two-week class on-site followed by an extended training period that can last up to a year. NLM would like to shorten the training time.

Each indexer produces about four articles per hour. The full-screen 3270 indexing system has many validations. For example, if the document mentions "testicular neoplasm", the system will automatically add "male" as the check tag. "Pregnancy" automatically adds "female". There are many warning messages.

MeSH is updated annually. When the update occurs, NLM often adds more specific terms. For example, acupuncture now has specific terms underneath it in the hierarchy. In this case, the system will provide a warning message if the general term is used by an indexer, so that the indexer will look to see if the specific term is applicable to the document.

The quality assurance unit does ongoing maintenance to the file including retrospective changes. One source of quality control for index terms comes from the millions of searches performed on the database each year. Users may locate incorrect indexing. The in-house staff can identify where the problem is and make corrections.

Data entry was done by a contractor for many years. NLM continues to keyboard the cataloging portion of the record, but it has been looking at alternatives. Currently, some material is captured via optical character recognition. In addition, NLM has an agreement with some publishers to obtain SGML (Standard Generalized Markup Language)-tagged bibliographic data. NLM has developed its own DTD (document type definition) to which most publishers adhere. A few variations of DTD are available for large publishers. NLM now reviews 600-700 articles per week in SGML format and the number continues to grow.

MEDLINE on the Web is now free. The MEDLINE system itself (including the Elhill search engine) is in the process of being phased out and replaced by the search engine used for the PUBMED system developed by the National Center for Biotechnology Information located at the Lister Hill Center. PUBMED allows the user to link to the Web-based full-text journal distributed from the publisher via the Internet.

The Department of Health and Human Services (DHHS) is trying to reduce the number of mainframe computers. The NLM mainframe must be eliminated; this is one of the reasons why the indexing and MEDLINE search system must change. The new indexing system will be implemented in a client/server environment. It will be available around the clock. Indexing online from international locations will then be more viable.

NLM is also investigating ways to increasingly produce its indexing terms automatically. Work on the NGI (Next Generation Indexing) has been going on for about a year. The NGI group meets monthly and conducts or oversees research projects. One experiment uses a meta-map project based on the Unified Medical Language System (UMLS) metathesaurus and text to

propose candidate terms. This is a small scale project where several indexers access and evaluate articles that have candidate indexing included, based on the meta-map. The results of the study indicate that half of the terms have to be deleted; thus, slowing down the indexers. (NASA reported a similar experience with the processing of candidate MAI terms. In this environment, the indexers begin to perform a completely separate task of analyzing and assessing the terms produced automatically.) The software program at NLM can still be improved to eliminate some of the noise. Testing will continue.

The NLM Web site has a wealth of information on indexing and MeSH. The NGI also has a web site.

2.6   National Air Intelligence Center (NAIC)

NAIC serves as the executive agent for DIA's Defense Intelligence Information Services Program (DIISP) of which the Central Information Reference and Control (CIRC) database is a part. CIRC currently holds well over ten million documents that serve the intelligence information needs of the three armed forces as well as other government or government-sponsored research and development agencies. The CIRC database offers a wide diversity of material ranging from foreign periodicals, patents, and brochures to finished intelligence studies and intelligence information reports.

Each CIRC document contains three parts—the bibliographic material, the text, and the indexed portion. Bibliographic information includes such elements as title, document number, country, dates of information, publication data, secondary source publication data, microfiche number, and COSATI subject code. Text may be entire or an abstract or extract. The uniqueness of CIRC among other government and commercial databases comes from the depth to which the information is indexed. This allows for more thorough and efficient retrieval to better serve the analytic needs of the intelligence community.

Because of in-depth indexing, every CIRC file identifies all personality names, organizations, facilities, equipment, or nomenclature. This PFN data is recorded along with the attributes and the relationships that exist between the people, facilities and equipment. When the user retrieves a CIRC record, it will specify that people are members of or related to specific organizations. If organizations or equipment are known by other names, that information is identified as well as the developers and designers of given nomenclature. Currently, a machine-aided indexing tool is being developed to more efficiently process information that automatically extracts this PFN by using natural language processing.

2.7   Defense Technical Information Center (DTIC)

DTIC's multidisciplinary Technical Report (TR) database contains over two million bibliographic records covering 25 broad subject fields and 251 subgroups ranging from aviation technology to communications. Recent statistics on the TR Database collection show that the top five subject fields in the database are Physics (over 226,000 records), Earth Science and Oceanography (165,000 records), Behavioral Science and Social Science (150,000 records), Behavioral and Social, Mathematical and Computer Science (144,00 records), and Biological and Medical Science (133,500 records).

DTIC utilizes both controlled and uncontrolled vocabulary for indexing. All technical reports except multimedia documents are processed electronically via the Electronic Document Management System (EDMS). The EDMS system was implemented in 1995. The Subject Analysis Branch underwent a reorganization in May of this year which combined the cataloging and indexing functions under one branch. Bibliographic analysts catalog, abstract, and index approximately 1400 documents biweekly. The full electronic document can be viewed on the computer screen. The system has a searchable electronic thesaurus online. The indexers are assigned documents according to their subject specialties. They access the document, enter the cataloging data, block the abstract text, convert the image to ASCII by using optical character recognition, click on a machine aided indexing button, and suggest indexing terms are posted in the citation window on the computer screen. The indexer reviews the MAI-suggested terms and adds additional thesaurus and non-thesaurus terms as needed. Index terms that cover the main topic of the document are asterisked or weighted. Subject category fields are assigned according to the subject content of the report. The document is then saved and sent to the Quality Assurance Branch for a final review of typographical errors before being sent to a validation system, and then to DROLS (Defense Research, Development, Test, and Evaluation Online System), DTIC's information retrieval system.

DTIC customers use index terms to retrieve relevant documents. In-house retrievers use asterisked terms and subject category fields to set up user profiles for searching.

The DTIC Thesaurus is updated on a quarterly basis and a new thesaurus is published every three years. DTIC is currently working on adding related terms to the thesaurus and securing LEXICO Thesaurus Maintenance software in order to better manage thesaurus maintenance.

The most common concerns DTIC has about indexing are how to maintain quality in a production environment and determining if full-text, search capability negates the need for human indexing.

3.0   SUMMARY

The main issues related to indexing shared by the CENDI agencies are as follows.

Software/technology identification for automatic support to indexing. As the resources for providing human indexing become more precious, agencies are looking for technology support. DTIC, NASA, and NAIC already have systems in place to supply candidate terms. New systems are under development and are being tested at NAIC and NLM. The aim of these systems is to decrease the burden of work borne by indexers.

Training and personnel issues related to combining cataloging and indexing functions. DTIC and NASA have combined the indexing and cataloging functions. This reduces the paper handling and the number of "stations" in the workflow. The need for a separate cataloging function decreases with the advent of EDMS systems and the scanning of documents with some automatic generation of cataloging information based on this scanning. However, the merger of these two diverse functions has been a challenge, particularly given the difference in skill level of the incumbents.

Thesaurus maintenance software. Thesaurus management software is key to the successful development and maintenance of controlled vocabularies. NASA has rewritten its system internally for a client/server environment. DTIC has replaced its systems with a commercial-off-the-shelf product. NTIS and USGS/BRD are interested in obtaining software that would support development of more structured vocabularies.

Linked or multi-domain thesauri. Both NTIS and USGS/BRD are interested in this approach. NTIS has been using separate thesauri for the main topics of the document. USGS/BRD is developing a controlled vocabulary to support metadata creation and searching but does not want to develop a vocabulary from scratch. In both cases, there is concern about the resources for development and maintenance of an agency-specific thesaurus. Being able to link to multiple thesauri that are maintained by their individual "owners" would reduce the investment and development time.

Full-text search engines and human indexing requirements. It is clear that the explosion of information on the web (both relevant web sites and web-published documents) cannot be indexed in the old way. There are not enough resources; yet, the chaos of the web bets for more subject organization. The view of current full-text search engines is that the users often miss relevant documents and retrieve a lot of "noise". The future of web searching is unclear and demands or requirements that it might place on indexing is unknown.

Quality Control in a production environment. As resources decrease and timeliness becomes more important, there are fewer resources available for quality control of the records. The aim is to build the quality in at the beginning, when the documents are being indexed, rather than add review cycles. However, it is difficult to maintain quality in this environment.

Training time. The agencies face indexer turnover and the need to produce at ever-increasing rates. Training time has been shortened over the years. There is a need to determine how to make shorter training periods more effective.

Indexing systems designed for new environments, especially distributed indexing. An alternative to centralized indexers is a more distributed environment that can take advantage of cottage labor and contract employees. However, this puts increasing demands on the indexing system. It must be remotely accessible, yet secure. It must provide equivalent levels of validation and up-front quality control.