Conf-751062--1

LBL-4605

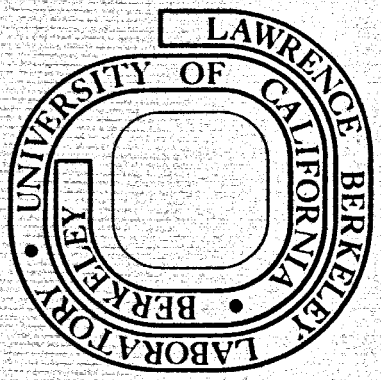
Presented at the 10th Annual Meeting of
the Geoscience Information Society,
Salt Lake City, UT, October 21, 1975

DEVELOPMENT OF A GEOTHERMAL THESAURUS

Jessie J. Herr

October 1975

Prepared for the U. S. Energy Research and
Development Administration under Contract W-7405-ENG-48



LBL-4605

MASTER

Development of a Geothermal Thesaurus*

Jessie J. Herr
Information Research Group
Lawrence Berkeley Laboratory
University of California
Berkeley, CA 94720

Introduction

A thesaurus of terminology associated with geothermal energy is being developed by the Lawrence Berkeley Laboratory's (LBL) Information Research Group. This project is being carried out in collaboration with two groups: LBL's Geothermal Information Group (known as GRID for historical reasons), which is preparing the National Geothermal Information Resource, a data compilation (Phillips 1975); and the Energy Research and Development Administration's (ERDA) Technical Information Center (TIC), which is doing the indexing for ERDA's Energy Data Base (Alexander 1975). The Geothermal Thesaurus is a thesaurus for use in indexing and retrieval; but, in addition, it contains information that permits its use as an interface between indexing systems.

One way of viewing an information-retrieval thesaurus is as an interface between its users and an information storage and retrieval system. In retrieval, a thesaurus serves to translate the query, originally in natural language, into the vocabulary actually used for describing documents in the system and to suggest alternative ways of describing the subject of the search within the context of the particular system. During indexing, a thesaurus is used to translate the vocabulary of each document into the regularized vocabulary of the system and to indicate the document's relationship to others in the collection.

A thesaurus for use in an information storage and retrieval system basically consists of a collection of terms along with information about those terms. The most important type of information contained is the relationships between the terms, which are used to thread through the thesaurus to find the most appropriate term to represent a concept. Other types of information included are text, such as scope notes, and codes, for instance, those representing a broad categorization of the terms.

The Geothermal Thesaurus is an information-retrieval thesaurus and contains the types of information required for its use for vocabulary control and assistance during subject indexing and retrieval. However, in the Geothermal Thesaurus the concept of a thesaurus has been broadened

*Work performed under the auspices of the U.S. Energy Research and Development Administration.

DISCLAIMER

beyond that of an interface between an indexer or searcher and the system governed by the thesaurus to include the role of interface between information systems. Its traditional role is carried out in large part through the network of links between terms within the geothermal vocabulary. Its role as interface between systems is based on inclusion in the thesaurus of links between the geothermal vocabulary and the vocabularies of other systems.

The linking of the geothermal vocabulary to other indexing vocabularies is a natural consequence of two inter-related aspects of the current status of data-base development. The first aspect is that of the increasing availability of bibliographic data bases in machine-readable form. The second is the recognition that there is overlap in the coverage of the various data bases and that, because they are in machine-readable form, there exists the possibility that material prepared for one data base with one set of users may be manipulated in order to satisfy the needs of another user population and be included in another data base.

The Geothermal Thesaurus contains links to a number of vocabularies; the most important and the most thoroughly developed links are those to the ERDA indexing vocabulary. These links are provided in anticipation of an active interchange of bibliographic data between the highly specialized National Geothermal Information Resource and the broadly based ERDA Energy Data Base. Because both data bases use the same indexing style, namely coordinate indexing based on a controlled vocabulary of indexing terms called descriptors, the major variance in indexing is associated with differences in the specificity of the vocabularies. These differences arise primarily from the facts that the National Geothermal Information Resource concentrates on geothermal energy and data, while ERDA's Energy Data Base is concerned with all aspects of energy sources and utilization and is bibliographic.

The development of the Geothermal Thesaurus will be discussed, beginning with an outline of its subject scope, sources and methods used in compiling the list of terms, and those parts of the thesaurus structure that are related to its use for vocabulary control and assistance during indexing and retrieval. Then, the techniques being used to link the Geothermal Thesaurus to other vocabularies will be described and some examples of their use given. Finally, thesaurus processing software developed at LBL that permits extension of the concept of a thesaurus will be briefly mentioned.

Subject Scope of the Geothermal Thesaurus

The subject scope of the Geothermal Thesaurus includes the characterization of geothermal resources, including geographic distribution and geological and hydrological properties; detection and evaluation of geothermal resources; extraction of geothermal fluids, including reservoir engineering, drilling techniques, and topics associated with the transport of the fluids, such as corrosion; utilization of geothermal resources for both electric-power generation and non-electrical uses, such

as district heating; legal, economic, and environmental aspects of geothermal utilization; and basic laboratory studies related to the characterization and utilization of geothermal resources.

Compiling the List of Terms

A variety of sources have been used in compiling the list of candidate terms. One group of sources has been used to ensure that the vocabulary reflects that of the literature. In this group are indexing from TIC and GRID, indexing resulting from the LBL translation of the indexing for a 14,000-item bibliography (Summers 1971), and titles, abstracts, and indexing of geothermal papers retrieved in searches of a number of machine-readable data bases. Previous vocabulary efforts (thesauri, glossaries, indexing vocabularies) have been used to place the geothermal vocabulary in a more general context; the most important of these is the INIS Thesaurus (INIS 1974) upon which the ERDA indexing vocabulary is based. Data formats developed for the International Geothermal Information Exchange Program (Clark 1975), GRID specifications of frameworks for the analysis and description of data, and reports on research needs are important guides for terms to be used in data description. Finally, numerous reference works, review articles, and discussions with scientists and engineers engaged in geothermal research have provided an overview of the subject and the general outline of the thesaurus.

The tendency has been to include, rather than exclude, terms of unknown usefulness, and to provide paths through the thesaurus to make these terms accessible. The inclusion of questionable terms is feasible because of the limited size of the thesaurus at this time, and is desirable because the terminology requirements for data description and storage are under investigation.

Although the inclusion of terms is fairly unrestricted, those that are entered into the thesaurus are regularized in form in accordance with standard thesaurus-development practice. The list of terms will be reviewed periodically. Those that are used infrequently will be eliminated from the thesaurus so they won't impede locating those terms that are used; heavily used terms will be subdivided, if possible, to ensure effective retrieval.

At present, the Geothermal Thesaurus contains nearly 1300 terms, with about 2300 relationships between terms, more than 600 lines of text, and 2300 links between terms and codes. Nearly all subjects within the scope are represented by terms of sufficient specificity for general subject indexing. More detailed terminology is available for those subjects upon which GRID has concentrated (geological setting of geothermal resources, geochemical and geophysical exploration, and thermodynamic properties of aqueous solutions of interest in geothermal systems).

The Structure of the Geothermal Thesaurus

The relationships that have been used in the Geothermal Thesaurus are the following (*indicates links used to relate the Geothermal Thesaures to other vocabularies):

Links between terms

USE	USE instead of (USE...AND...)
UF, UF+	Used For, Used For in combination
SEE	See (SEE...OR...)
SF	Seen From
BT	Broader (more general) Term
NT	Narrower (more specific) Term
RT	Related Term
*GT	Goes To (translates to)
*CF	Comes From (translates from)

Links between terms and text

SN	Scope Note
DF	Definition

Links between terms and codes

SC	Subject Category
*SO	Source of the term

Information about Terms within the Geothermal Vocabulary

USE references lead the thesaurus user from a term that is not acceptable in the system to one that is. Often the terms are synonyms; to avoid dispersing information in the system, one of the pair is chosen as the term to be used. For instance, ALUMSTONE and ALUNITE are synonyms for a mineral of interest in geothermal systems and the Geothermal Thesaurus contains the entry,

ALUMSTONE
  USE ALUNITE

which is read as, ALUNITE is to be used instead of ALUMSTONE. All term–term relationships are reciprocated, and the entry for ALUNITE includes the reciprocal of the ALUMSTONE entry,

ALUNITE
  UF ALUMSTONE

The USE reference is also used in connection with abbreviations, for instance,

UNITED STATES OF AMERICA
  USE USA

USA
  UF UNITED STATES OF AMERICA

Occasionally, a term is designated as unacceptable for use and the concept is to be represented by a combination of terms. In such a case, the reciprocal UF+ is used to indicate that the term is used in combination with other terms as a replacement for an unacceptable term.

The SEE reference refers from an unacceptable term to more than one alternative terms, any one of which would be acceptable, for instance,

CONDUCTIVITY
  SEE ELECTRIC CONDUCTIVITY
  OR THERMAL CONDUCTIVITY

with the reciprocals

ELECTRIC CONDUCTIVITY
  SF CONDUCTIVITY

THERMAL CONDUCTIVITY
  SF CONDUCTIVITY

The Broader Term and Narrower Term relationships show class membership or geographic inclusion. A BT relationship leads the user to a more general term and an NT leads to a more specific term, for instance,

OFFICE BUILDINGS	COMMERCIAL BUILDINGS
BT COMMERCIAL BUILDINGS	NT OFFICE BUILDINGS

In addition, the Thesaurus displays all of the broader or narrower terms for a particular term, this is called display of the term hierarchy. Continuing the previous example, COMMERCIAL BUILDINGS has the broader term BUILDINGS (the number following the relationship label indicates the level in the hierarchy),

COMMERCIAL BUILDINGS
BT1 BUILDINGS
NT1 OFFICE BUILDINGS

which leads to the fully developed BT hierarchy for OFFICE BUILDINGS and the NT hierarchy for BUILDINGS,

OFFICE BUILDINGS	BUILDINGS
BT1 COMMERCIAL BUILDINGS	NT1 COMMERCIAL BUILDINGS
BT2 BUILDINGS	NT2 OFFICE BUILDINGS

The BT relationships assist the user in broadening a search and the NT's help in narrowing a search, or making it more specific.

The Related Term relationship is used to link terms that are not related by class membership, but indicating their existence may be helpful to the user in some way, for instance,

SALINITY	BRINES
RT BRINES	RT SALINITY

Terms that are sometimes, but not always, related by class membership are entered as Related Terms, instead of as a BT/NT pair.

Although the meaning and intended use of a term usually are defined by the terms linked to it, occasionally, further information will assist in the effective use of the thesaurus. This information is contained in text called scope notes, which are used to make the scope of use of a term absolutely clear.

As a normal part of thesaurus development, definitions have been obtained for many of the terms included in the thesaurus. Whenever possible, these definitions have been stored as a part of the information associated with the terms. Not all terms have definitions; however, an attempt has been made to include definitions for terms that are particularly important in geothermal energy or whose definitions are not readily available. Definitions are included as a special type of text (rather than as scope notes), so that they can be selectively printed.

Terms are assigned to one or more subject categories, which are represented by four-character codes. These codes are used to list the terms by subject ctaegory to provide an alternative to the alphabetical thesaurus as a means for finding the most appropriate terms. In the Geothermal Thesaurus, three types of subject categories have been used: (1) mission-oriented categories, such as geothermal exploration; (2) discipline-oriented categories, such as chemistry; and (3) categories representing long lists of similar terms, for instance, minerals.

Links between the Geothermal Thesaurus and Other Vocabularies

All of the relationships and other types of information described above are associated with the use of the Geothermal Thesaurus as an interface between an indexer or searcher and the data base. Next, we turn to those parts of the Thesaurus related to its use as an interface between its vocabulary and that of other systems. The term–code link Source and the term–term links Goes To and Comes From are used to tie the Geothermal Thesaurus to other vocabularies. The Source code identifies the vocabulary to which a term belongs. At present, the most significant source code is that which indicates that the term is a part of the ERDA indexing vocabulary. Other source codes are being included in anticipation of the use of the thesaurus to assist in screening machine-readable data bases for geothermal-related material.

The Goes To and Comes From links are used in conjunction with the source code indicating the ERDA indexing vocabulary. They will be used to translate the highly specialized vocabulary being used by GRID into the more general vocabulary used for ERDA indexing. Similarly, they will be used to translate ERDA indexing into the GRID vocabulary; however, only a partial translation will be possible because it will be from the general to the specific.

There are three general situations in which these inter-vocabulary links are being used. In the first, the vocabulary for GRID is more specific than that for ERDA. As an example, the ERDA terminology includes GEOTHERMOMETERS, while that for GRID also includes two specific types of geothermometers, GEOCHEMICAL THERMOMETERS and ISOTOPE GEOTHERMOMETERS. In this case, the Thesaurus contains the following entries:

GEOCHEMICAL THERMOMETERS
  SO GRID
  GT GEOTHERMOMETERS

ISOTOPE GEOTHERMOMETERS
  SO GRID
  GT GEOTHERMOMETERS

with the reciprocals,

GEOTHERMOMETERS
  SO ERDA GRID
  CF GEOCHEMICAL THERMOMETERS
  CF ISOTOPE GEOTHERMOMETERS

With these entries, if an item indexed with GEOCHEMICAL THERMOMETERS by GRID is to be included in ERDA's Energy Data Base, this term will first be replaced by GEOTHERMOMETERS to make the indexing compatible with the rest of the Energy Data Base.

Another case in which the Goes To relationship is utilized is that of terms that are unambiguous within the limited context of geothermal energy, but which would be ambiguous in the broader context of the Energy Data Base. Such terms are qualified in the ERDA vocabulary to make their meanings quite explicit; however, this qualification is unnecessary and artificial within the geothermal-only scope. An example is the term FAULTS, which appears in the ERDA vocabulary as GEOLOGIC FAULTS. To permit each to use the most appropriate form, while maintaining compatibility, the Geothermal Thesaurus contains

FAULTS	GEOLOGIC FAULTS
SO GRID	SO ERDA
GT GEOLOGIC FAULTS	CF FAULTS

Finally, the Goes To relationship is applied to cases in which there are differences in the meaning of terms, some of which are common to both vocabularies. For instance, both systems define a series of pressure ranges. The ERDA vocabulary defines HIGH PRESSURE to cover the range from 100 to 1000 atmospheres, while GRID subdivides the range into ELEVATED PRESSURE, for 100 to 500 atmospheres, and HIGH PRESSURE, for pressure greater than 500 atmospheres. To provide compatibility, the Thesaurus includes the entries,

ELEVATED PRESSURE	HIGH PRESSURE
SO GRID	SO ERDA GRID
GT HIGH PRESSURE	CF ELEVATED PRESSURE

An item indexed by GRID with the term HIGH PRESSURE would not be modified; however, the term ELEVATED PRESSURE would be transformed to HIGH PRESSURE before input to the ERDA system.

Thesaurus Processing Software

The Geothermal Thesaurus is being processed using software developed at LBL. The major purpose of this software, as with most thesaurus processing systems, is to reduce the clerical work associated with thesaurus building. The software performs tasks such as: checking the spelling of terms being entered into the thesaurus; reciprocating all term–term and term–code links; checking the consistency of term–term links; and generating the hierarchies for BT and NT relationships.

In most thesaurus processing systems, the types of relationships are embedded in the program, that is, the program specifies the types of relationships that can be used. In LBL's system, the thesaurus builder specifies the structure, or types of relationships, to be used and this may be different for every thesaurus. This feature was included to permit the addition of nonstandard relationships, such as Goes To and Comes From, and to permit the processing of thesauri developed using other software packages, each of which defines a different set of relationships.

A thesaurus publishing program is under development to permit the flexible formatting of the printed thesaurus and the selection of the information to be printed. With it, if desired,

definitions could be printed in an auxiliary publication rather than in the thesaurus proper, and information, such as term sources and the vocabulary-translating relationships, which is not of interest to most thesaurus users, could be printed only in the thesaurus editor's master copy of the thesaurus.

Summary

At present, the Geothermal Thesaurus contains 1300 terms. Nearly all subjects within the scope are represented at a specificity level sufficient for general subject indexing; some topics are covered in the detail required for the description of data and for capturing text-type data, such as the geological setting of a geothermal field. The Geothermal Thesaurus contains information, usually associated with information-retrieval thesauri, which is designed to assist its users in finding the most appropriate term(s) to represent a concept. In addition, it contains information that relates its vocabulary to other vocabularies. In particular, it has links to the ERDA indexing vocabulary, which will be used to provide compatibility between a highly specialized, data-oriented geothermal-energy information system and a broadly based, bibliographic system concerned with all aspects of energy. Thesaurus software developed at LBL performs the normal processing functions and permits flexible definition of the structure of a thesaurus.

References

C. G. Alexander, "ERDA Information Programs," to be presented at the Tenth Annual Meeting of the Geoscience Information Society, Salt Lake City, 21 October 1975.

A. L. Clark, J. A. Calkins, E. Tongiorgi, and E. Stefanelli, "A Report on the International Geothermal Information Exchange Program" presented at the Second United Nations Symposium on the Development and Utilization of Geothermal Resources, San Francisco, May 1975.

INIS Section, International Atomic Energy Agency, INIS: Thesaurus, IAEA-INIS-13 (Rev. 7), July 1974.

S. L. Phillips, "The National Geothermal Information Resource," to be presented at the Tenth Annual Meeting of the Geoscience Information Society, Salt Lake City, 21 October 1975.

W. K. Summers (comp.), Annotated and Indexed Bibliography of Geothermal Phenomena, New Mexico State Bureau of Mines and Mineral Resources, Socorro, 1971.

TECHNICAL INFORMATION DIVISION
LAWRENCE BERKELEY LABORATORY
UNIVERSITY OF CALIFORNIA
BERKELEY, CALIFORNIA 94720