SAF/AQT-SR-90-007

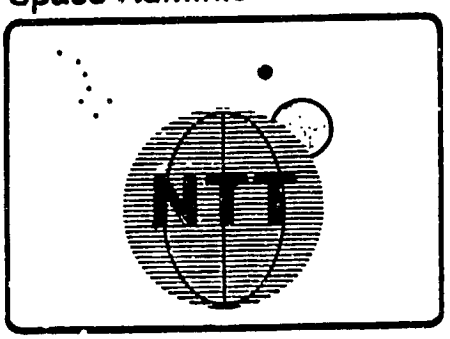
THE NATO THESAURUS PROJECT

Jonathan Krueger

Control Data Corporation
Alexandria, Virginia

May 1990

NASA
National Aeronautics and Space Administration



UNITED STATES AIR FORCE
SCIENTIFIC AND TECHNICAL INFORMATION PROGRAM
CONTRIBUTIONS TO INFORMATION SCIENCE

USAF STINFO CONTRIBUTION 90/3
JOINT REPORT

SECRETARY OF THE AIR FORCE
DEPUTY FOR SCIENTIFIC AND
TECHNICAL INFORMATION
(SAF/AQT) THE PENTAGON
WASHINGTON, DC 20330-1000

NATIONAL AERONAUTICS AND SPACE
ADMINISTRATION
SCIENTIFIC AND TECHNICAL INFORMATION
DIVISION (NTT)
WASHINGTON, DC 20546

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	May 90	

4. TITLE AND SUBTITLE
The NATO Thesaurus Project

5. FUNDING NUMBERS
C MDA903-88-C-0186

6. AUTHOR(S)
Jonathan Krueger

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Control Data Corp.
4900 Seminary Rd.
Alexandria, VA 22311

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Secretary of the Air Force
Deputy for Scientific and Technical Information
SAF/AQT, The Pentagon
Washington, DC 20330-1000

10. SPONSORING/MONITORING AGENCY REPORT NUMBER
USAF-STINFO
Contribution 90/3

SAF/AQT SR-90-007

11. SUPPLEMENTARY NOTES
Jointly sponsored by National Aeronautics and Space Administration, Scientific and Technical Information Div. (NTT), Washington, DC 20546.
USAF STINFO Program, Contributions to Inormation Science.

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release;
distribution unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

This document describes functionality to be developed to support the NATO technical thesaurus. Described are the specificity of the thesaurus structure and function; the distinction between the thesaurus information and its representation in a given online, machine readable, or printed form; the enhancement of the thesaurus with the assignment of COSATI codes (fields and groups) to posting terms, the integration of DTIC DRIT and NASA thesauri related terminology, translation of posting terms into French; and the provision of a basis for system design.

14. SUBJECT TERMS
Thesauri, Thesauri integration; Bilingual thesauri, Vocabulary; DRIT, DRIT terminology, NASA, NASA terminology, English terminology; French terminology; NATO, DTIC

15. NUMBER OF PAGES
33

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UL

THE NATO THESAURUS PROJECT

Jonathan Krueger

Control Data Corporation

ABSTRACT

This document describes functionality to be developed to support the NATO technical thesaurus. The intended audience includes project sponsors, users, developers, and maintainers. No special technical knowledge is assumed.

Acknowledgements: the assistance and cooperation of Charles R. Jacobs and Ronald L. Buchan, chief lexicographers of DTIC and NASA respectively, and the time and expertise of June P. Silvester and Charles Brown, NASA STIF linguistic and data processing staff members respectively, are gratefully acknowledged.

SECTION 1: GENERAL

1.1 Purpose

This functional discription of the NATO technical thesaurus is written to:

a. specify the structure and function of the thesaurus

b. distinguish between the information contained in the thesaurus and its representation in a given online, machine readable, or printed form

c. explain how the thesaurus will be enhanced with

- assignment of COSATI codes (fields and groups) to posting terms
- integration of NASA related terminology
- translation of posting terms into French

d. provide a basis for system design

1.2 Project References

The general nature of the programs to be developed is information retrieval, selective dissemination of information, database management, and information resources management.

Relevant references include:

a. An Operational System for Subject Switching Between Controlled Vocabularies: A Computational Linguistics Approach. June P. Silvester, Roxanne Newton, and Paul Klingbiel. NASA Contractor Report 3838, October 1984.

b. DTIC Retrieval and Indexing Terminology, 3rd Ed. Defense Technical Information Center. January 1987. AD-A176 000.

c. NASA Thesaurus, Vols. 1, 2, 3. National Aeronautics and Space Administration, Scientific and Technical Information Division, 1988. NASA SP-7064.

d. Subject Categorization Guide for Defense Science and Technology, Defense Technical Information Center. October 1986. AD-A172 650.

e. Thesaurus of Thesaurus Terms (TOTT). ASTM Committee on Terminology Working Group 6 Thesauri, March 1990.

1.3 Terms and Abbreviations

Term	Definition
COSATI	Committee on Scientific and Technical Information
DRIT	DTIC Retrieval and Indexing Terminology
DTIC	Defense Technical Information Center.
NASA	National Aeronautics and Space Administration; in this document, NASA will in context refer instead to the NASA technical thesaurus
NATO	North Atlantic Treaty Organization
STIF	Scientific and Technical Information Facility (a NASA facility)

SECTION 2: DRIT

2

The NATO thesaurus will be based on the DRIT. This section describes the structure and existing functionality of the DRIT.

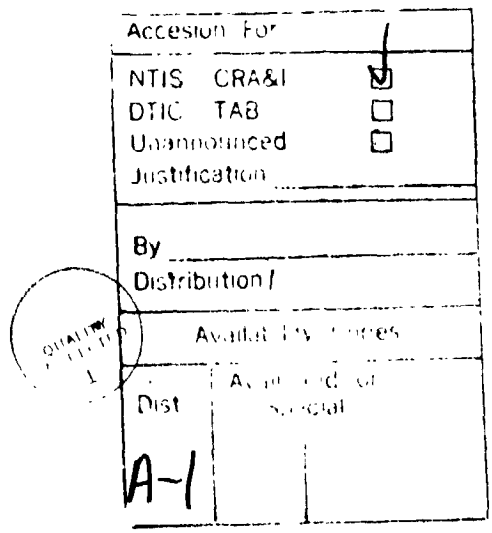
2.1 Lexical format of DRIT descriptors

The DRIT (reference 1.2.b) is a technical thesaurus. DRIT terms are single or multiple word descriptors. Since case distinctions are not significant, DRIT terms are usually shown in all upper case. Four examples are TOPOGRAPHY, CLEANING COMPOUNDS, CLEARANCES, and DETERGENTS. The character set is the mono-case alphabetic characters A through Z, the parentheses "(" and ")", the dash "-", the apostrophe "'", and the space used to separate words. [It's probably also true that terms cannot begin with a numeric digit; none currently do.] Four more examples are ANCHORS(MARINE), ANCHORS(STRUCTURAL), SELF-LOCKING NUTS, and BELLMAN'S INEQUALITY.

DRIT terms can be quite long. The longest term is currently NUCLEAR BIOLOGICAL CHEMICAL COLLECTIVE DEFENSE at 46 characters. The runners-up are VENEZUELAN EQUINE ENCEPHALOMYELITIS VIRUS and SEA BASED BALLISTIC MISSILE INTERCEPT SYSTEM at 44 and 46 characters respectively. Most terms are much shorter, however. The median length is 14 characters. Multiple word terms are slightly more common than single word terms.

2.2 Integrity constraints of DRIT descriptors

DRIT terms are unique: no two DRIT terms can be the same. There is no particular ordering or arrangement. They may be sorted alphabetically for convenience of display or use, but such ordering is not inherent. Therefore, they satisfy the definitional requirements to form a mathematical set. Sets have useful properties such as set union, set intersection, set difference, and subsets. Sets also have a concise and useful notation: the Venn diagram. Figure 1 is a simple Venn diagram that displays graphically most of the important information presented so far.



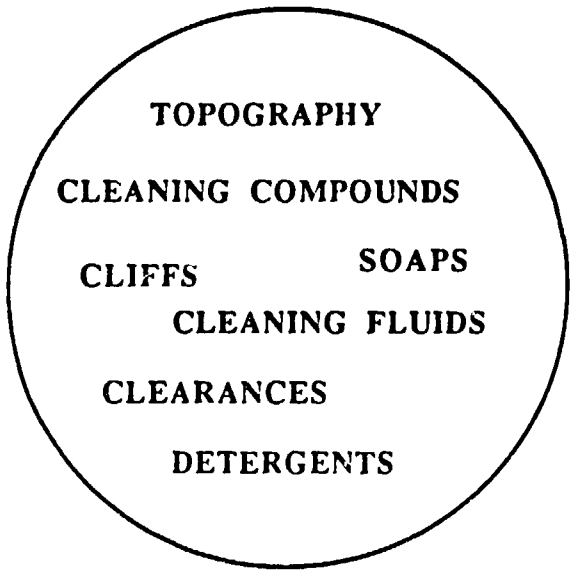


Figure 1. A set of descriptors

2.3 Relationships between descriptors

Within the DRIT distinctions are made between different types of terms and among their relationships with each other. Much of the existing structure is oriented toward use in assigning subject descriptors to technical reports, work groups summaries, and similar documents that DTIC serves as a repository for.

2.3.1 Posting Terms and Use References

Terms in the DRIT are divided into two distinct groups: posting terms and use references. Posting terms comprise the controlled vocabulary used to index DTIC scientific and technical information. Use references are terms that refer to a corresponding preferred posting term. For example, CLEANING FLUIDS is a use reference that refers to the posting term CLEANING COMPOUNDS. This is presented more concisely by the Venn diagram in Figure 2; posting terms and use references are simply two non-intersecting sets in a universe of descriptors.

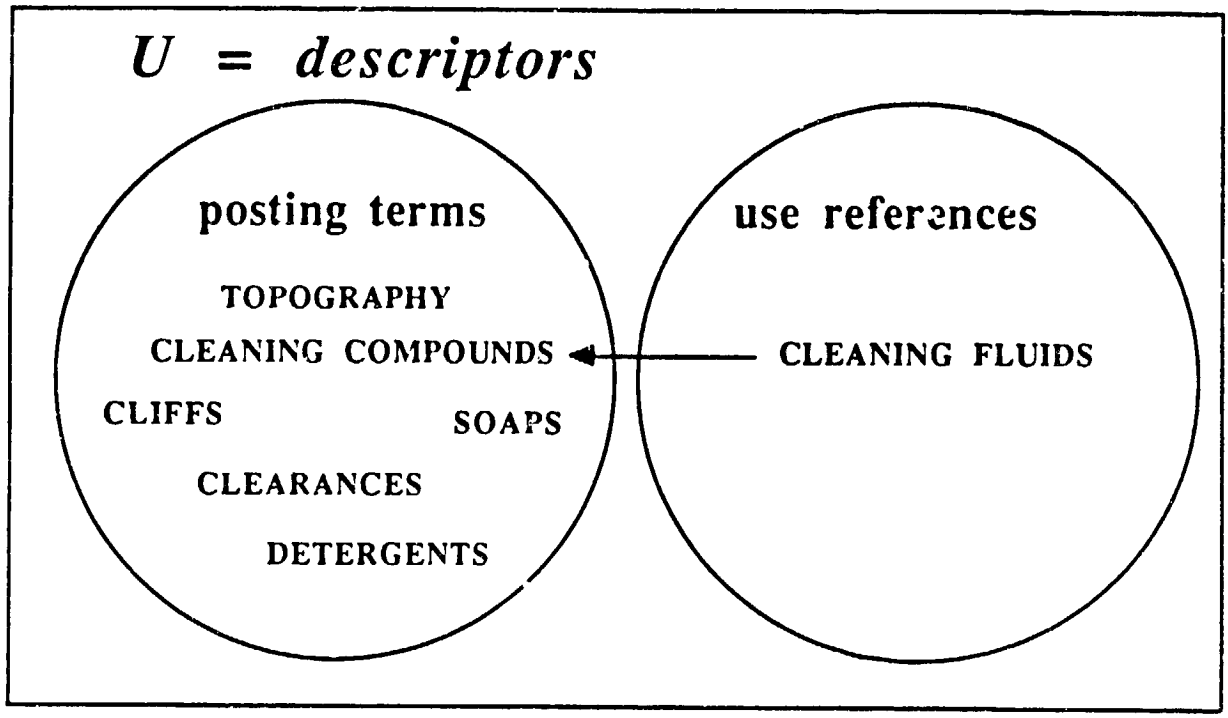


Figure 2. Two non-intersecting sets of descriptors

As the Venn diagram implies, the universe of descriptors may be larger than the union of the two sets shown. This is in fact the case if such descriptors as DTIC open-ended terms are included. For the purposes of this document, their inclusion is not in any way critical. Therefore, the universe is exactly the size of the union of the two sets shown. In the 1989 edition of the DRIT, there are 13430 posting terms, 2294 use references, for a total of 15724 terms.

2.3.2 Broader and Narrower Terms

Posting terms are further organized into broader and narrower terms. For example, the posting term TOPOGRAPHY has a narrower term CLIFFS. Broader and narrower terms reciprocate; TOPOGRAPHY is a broader term for CLIFFS. All broader and narrower terms in DRIT are reciprocal relationships between posting terms. Two different ways of diagramming this are shown in Figures 3 and 4.

Figure 3 shows the Venn circle denoting the set of posting terms. Some pairs of terms in this set are shown connected with a line that is broader on one end. This denotes a broader/narrower relationship between the two terms.

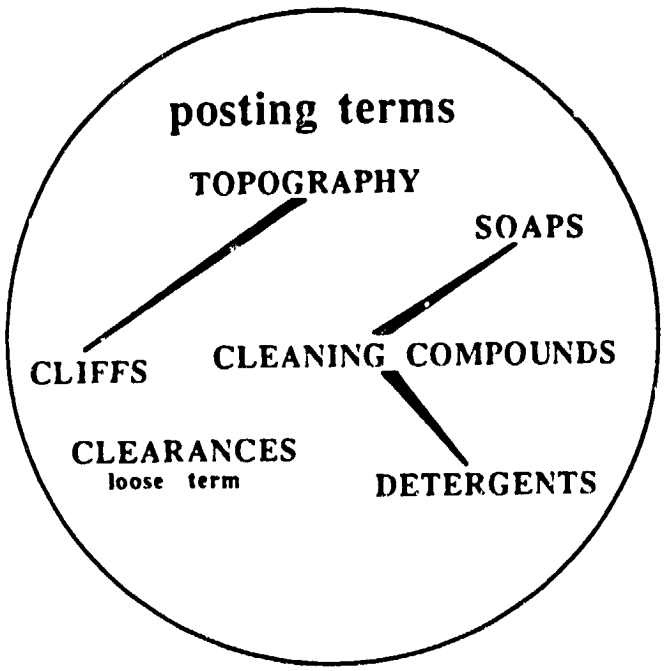


Figure 3. Relationships between posting terms

Figure 3 shows that some terms have relationships with multiple other terms. CLEANING COMPOUNDS is a broader term for SOAPS; CLEANING COMPOUNDS is also a broader term for DETERGENTS. Due to reciprocation, this also means that some terms have more than one narrower term. Some terms have no broader/narrower relationships with any other term. This simply means that no broader or narrower term is listed for these terms. They are called loose terms. The example shown is CLEARANCES.

2.3.3 Hierarchy

Since some terms can have more than one narrower term, an hierarchy is implied. Figure 4 diagrams an example:

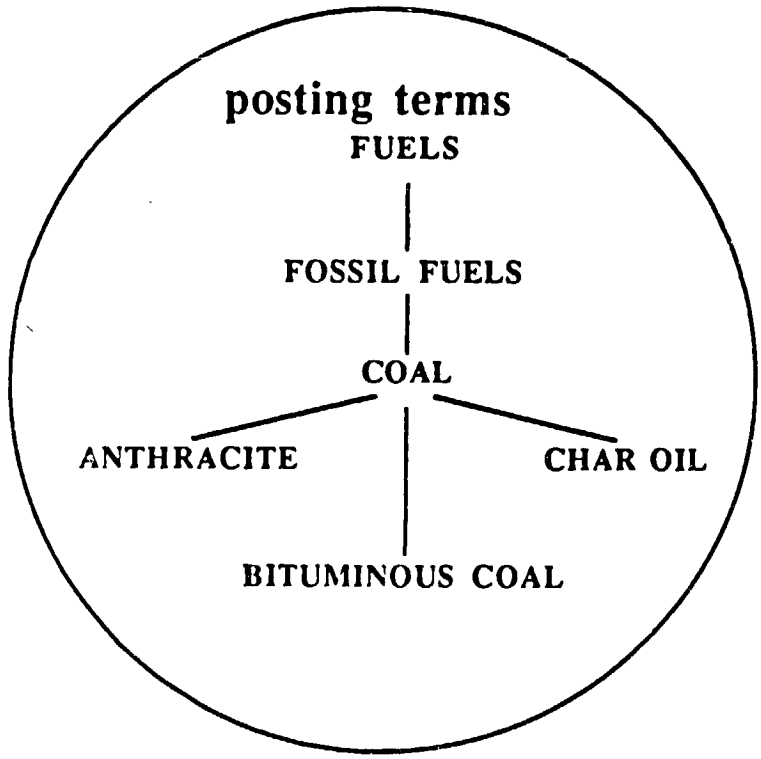


Figure 4. An hierarchy of posting terms

Figure 4 shows that with assignment of broader/narrower terms, the hierarchy can quickly form an extremely useful extended set of relationships. By using the hierarchy, posting terms can be chosen more specifically or more generically as required. In DRIT extensive use has been made of hierarchy. The example shown is 4 levels deep; hierarchies of 5 levels or more are common; the deepest hierarchy is ORGANIZATIONS with 10 levels.

Figure 4 shows FOSSIL FUELS as a narrower term for FUELS, and COAL as a narrower term for FOSSIL FUELS. Thus COAL is an even narrower term for FUELS; this demonstrates transitivity. To preserve transitivity, the still narrower term should not also be listed as a narrower term. For instance, COAL itself should not be listed as a narrower term for FUELS. It is the intent of DRIT to prevent this. Exceptions exist: CHEMICAL COMPOUNDS has narrower term INORGANIC COMPOUNDS; INORGANIC COMPOUNDS has narrower term BORATES; CHEMICAL COMPOUNDS has narrower term BORATES. Usually, however, transitivity of broader/narrower terms is preserved.

The transitivity is strictly ordinal, however. For example, the "amount narrower" between COAL and CHAR OIL has nothing to do with "amount narrower" between FOSSIL FUELS and COAL. It cannot be concluded that the "narrowing" for FOSSIL FUELS is the same as the "narrowing" for COAL. Furthermore, both ANTHRACITE and CHAR OIL are narrower terms for COAL, but it can not be concluded that they subdivide COAL equally. ANTHRACITE might be a much more specific kind of COAL than is CHAR OIL; the hierarchy does not say. More generally,

the top terms -- posting terms like FUELS in this example, with narrower but no broader terms -- are not equally broad, and are not intended to divide up the DRIT by subject area.

2.3.4 Polyhierarchy

As has been seen, as when some terms can have more than one narrower term, an hierarchy is implied. Similarly, when some terms can have more than one broader term, a polyhierarchy is implied. Figure 5 diagrams an example.

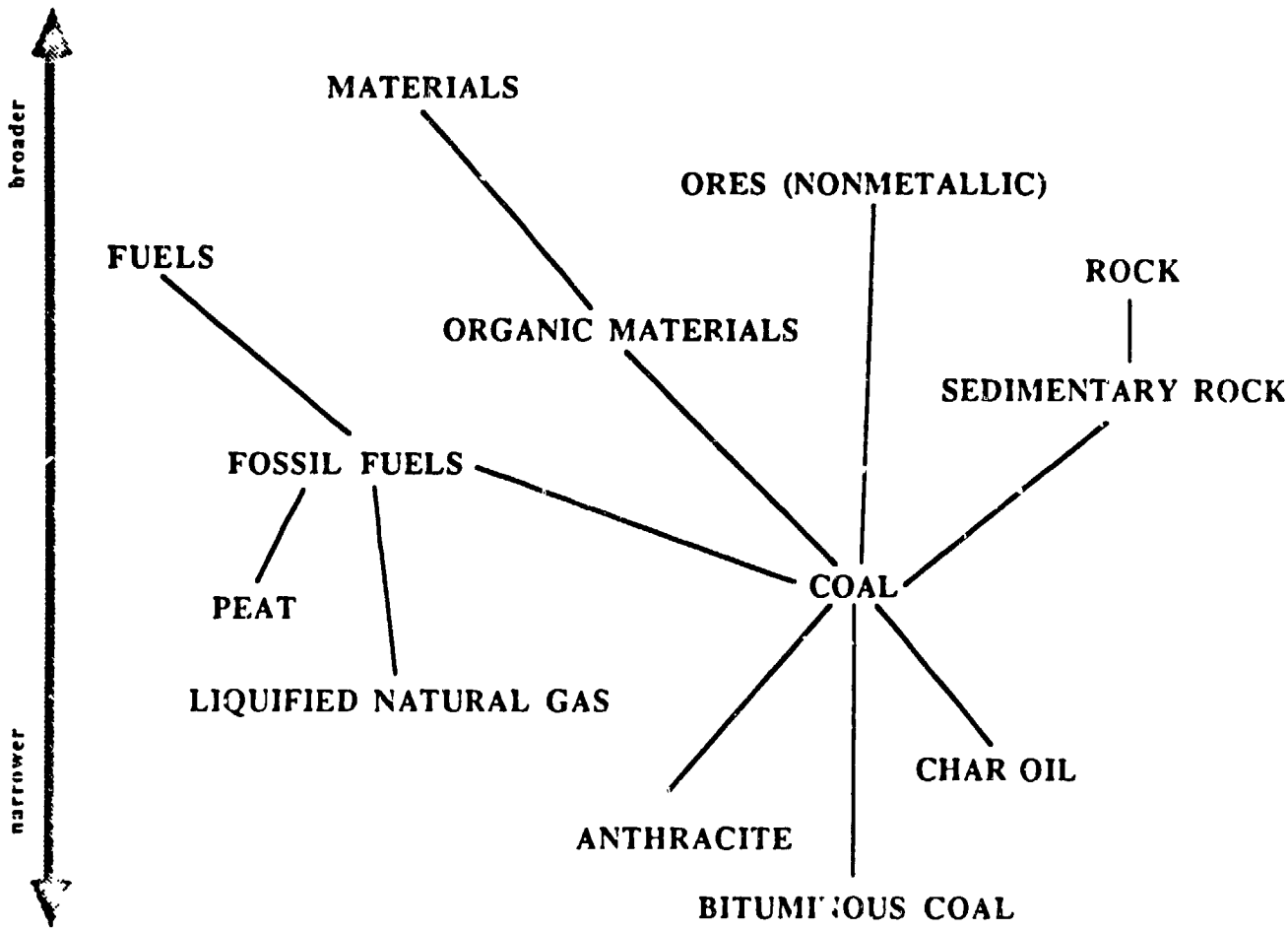


Figure 5. Polyhierarchy

Figure 5 shows that polyhierarchy make possible many more extended relationships. COAL has broader terms FOSSIL FUELS, ORGANIC MATERIALS, ORES(NONMETALLIC), and SEDIMENTARY ROCK, and narrower terms ANTHRACITE, BITUMINOUS COAL, and CHAR OIL. Therefore ANTHRACITE, for instance, has four top terms: FUELS, MATERIALS, ORES(NONMETALLIC), and ROCK. ANTHRACITE will appear in all four corresponding hierarchies. It will be in a different place in each hierarchy.

This reinforces the notion that transitivity is strictly ordinal and that top terms are not equally broad and not intended to be. If the opposite were true, it would be an anomaly that ANTHRACITE is two levels down from the top term ORES(NONMETALLIC) but three levels down from FUELS. This is not an anomaly. Nothing more is implied than that COAL is a broader term for ANTHROCITE, ORES(NONMETALLIC) is a broader term for COAL, FOSSIL FUELS is also a broader term for COAL, and FUELS is a broader term for FOSSIL FUELS.

Figure 4 shows the hierarchy as it is traditionally drawn: an inverted tree with its roots (top terms) at the top. Figure 5 continues use of the inverted tree to display

polyhierarchy. Figure 6 shows the same polyhierarchy but draws it with the roots at the bottom.

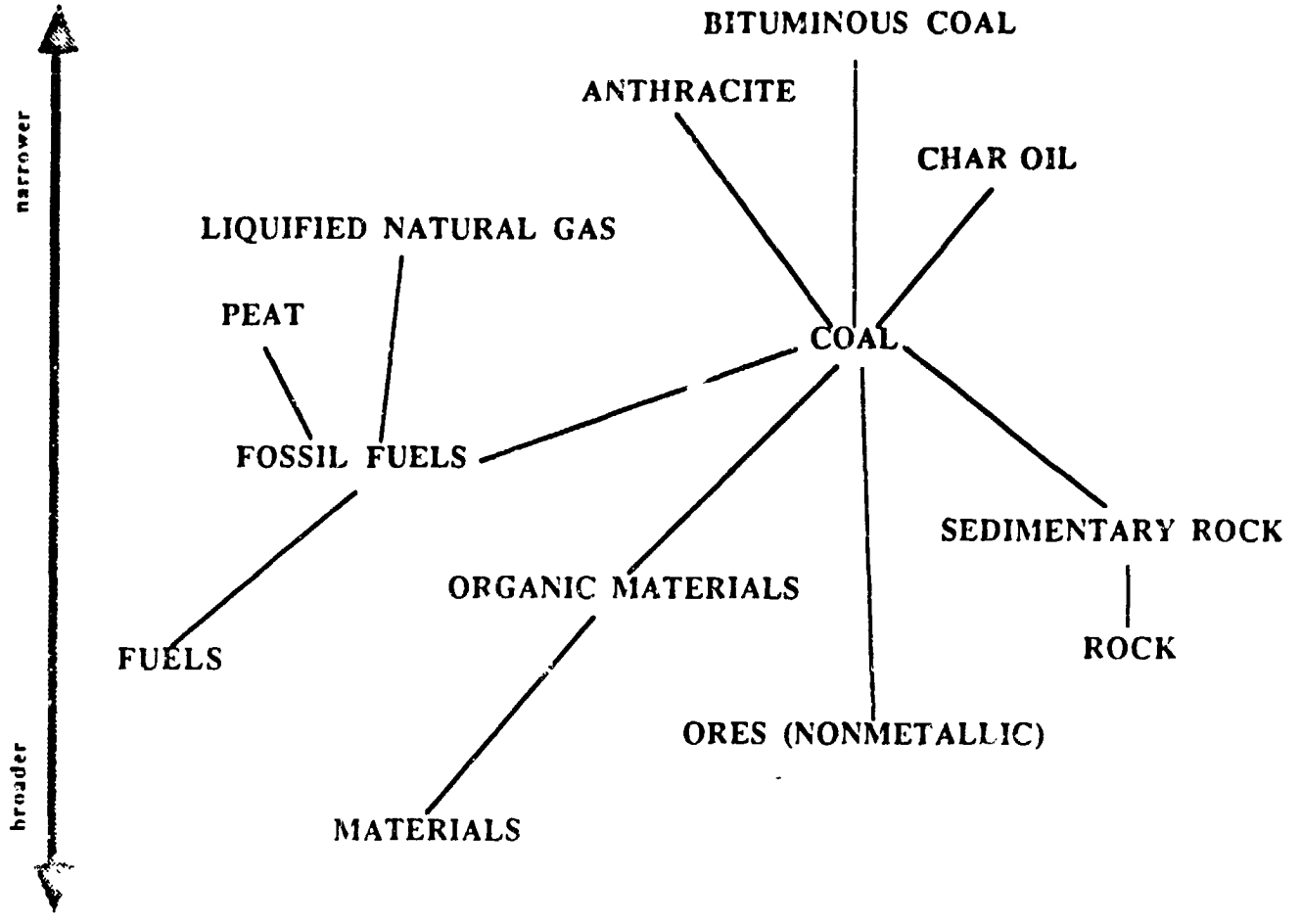


Figure 6. Polyhierarchy reconsidered: symmetry

What is revealed is symmetry: inversion of the broader/narrower axis does not change the information displayed. This implies an argument for polyhierarchy; it's a more flexible way to maintain relationships between terms. If no term were allowed to have more than one broader term, not only would one of the four broader terms for COAL have to be chosen, but COAL would also be limited to one of the four hierarchies. This in turn would mean that such terms as ANTHRACITE would be limited to the same hierarchy. Several extended relationships would become impossible. For instance, COAL is part of the FUELS hierarchy and the ROCK hierarchy; this is significant and would not be possible without polyhierarchy.

2.4 Use References and Use Combinations

Use references are further subdivided into use references and use combinations. Use references refer from a single deprecated term to a single preferred posting term. They reciprocate with "used for", commonly listed in thesaurus format as UF. Use combinations refer from a single deprecated term to multiple preferred posting terms to be used in combination. They reciprocate with "used for combination", commonly listed

in thesaurus format as UFC.

It can happen that multiple different use references refer to the same posting term. This is diagrammed in Figure 7.

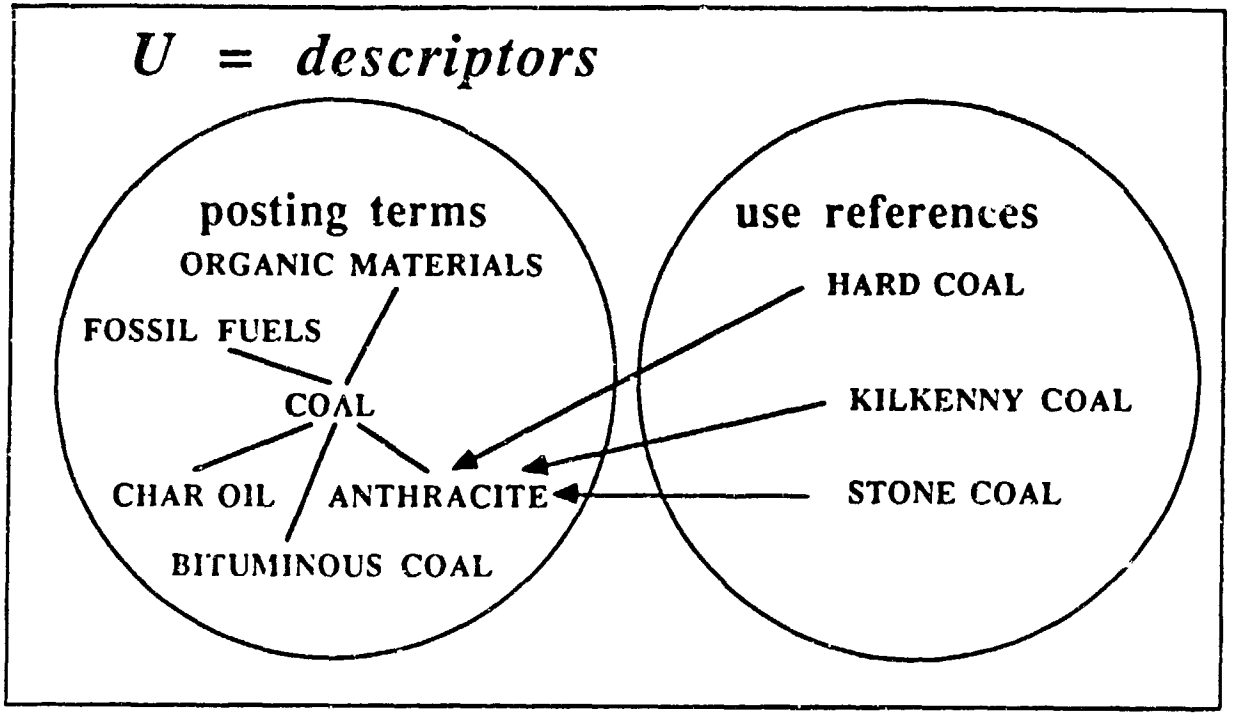


Figure 7. Multiple use references

Figure 7 shows that use references function as synonyms that point to a single term in the controlled vocabulary. Thus multiple use references are a natural outcome of standardizing on a posting term that replaces several terms from uncontrolled vocabulary. Each use reference refers from a single deprecated term to a single preferred term. Multiple use references happen to refer to the same single preferred term. This is *not* the same as use combination. Each use combination refers from a single deprecated term to several preferred terms to be used in combination. This is diagrammed in Figure 8.

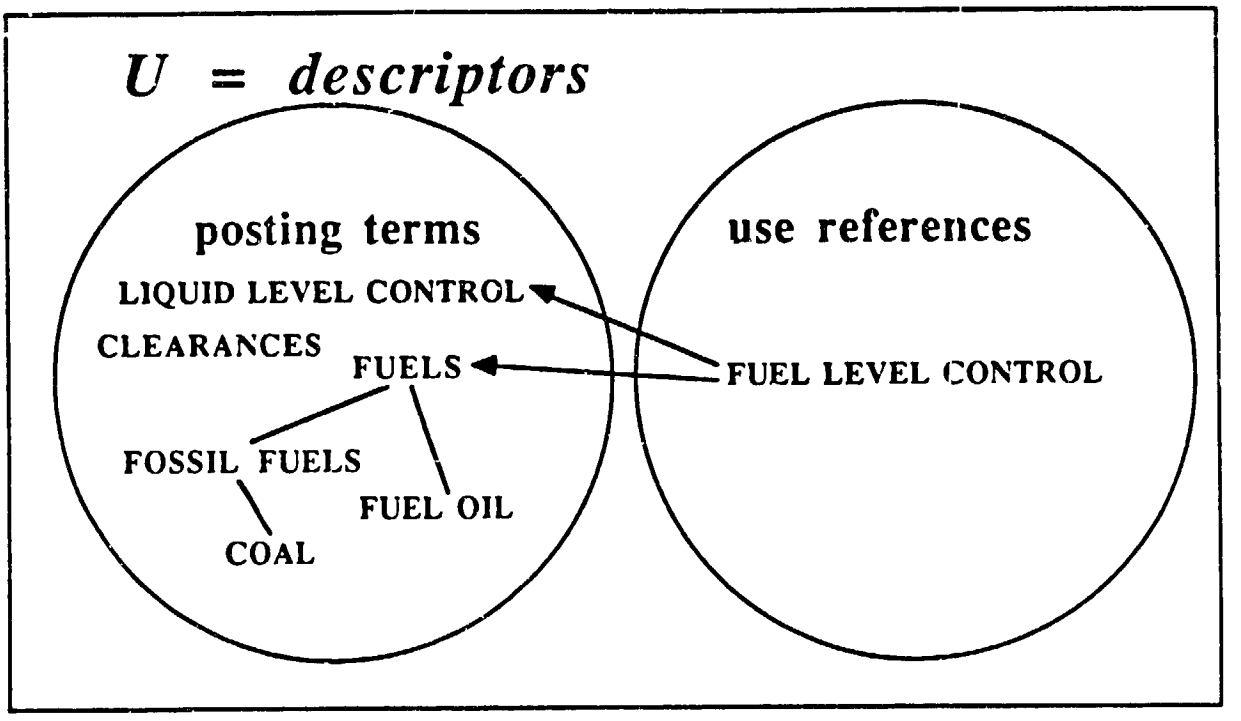


Figure 8. Use combination

The 1989 edition of the DRIT contains a total of 1934 use references and 360 use combinations.

Due to a limitation in vendor software, use combination and its reciprocal, used for combination, are not planned for inclusion in the NATO thesaurus.

2.5 Scope notes

DRIT posting terms can have scope notes, but most do not. Only about 1 in 20 terms has a scope note. For the more recently added terms, the scope note contains the date the term was added. Scope notes sometimes also serve as a source of extraneous information about the term; see below (section 2.6) for an example.

2.6 Subject codes

In the past, the DRIT has been maintained in parallel with "fields and groups". These are modified COSATI codes and are often referred to simply as COSATI codes (and sometimes also as fields/groups). They consist of a two-digit code for the field, representing one of 25 general fields of study, an optional (but usually present) two-digit code for the group, representing a group within that field of study, and an optional (and usually absent) two-digit code for the subgroup, representing a subgroup within that group. For instance, a field code of 11 represents Materials, and a group of 5 within field 11 represents Textiles. Thus a COSATI code of 11/05 indicates a subject area of Textiles (see Reference 1.2.d).

DTIC's maintenance of fields and groups has been applied principally to the administration of "need to know". Sensitive, confidential, and classified data are made available to cleared personnel based on need to know. Need to know may be established within a given subgroup, group, or field. This application suggests another merit of the codes: unlike broader/narrower terms, they are designed to divide areas of knowledge roughly equally. Thus the divisions of the 25 fields are of roughly comparable size, and the subdivisions of groups and subgroups are roughly equally large both within and across fields.

Since DTIC has maintained DRIT separately from fields and groups, the printed edition of the DRIT has not always listed them. The subject categorization guide (Reference 1.2.d) has listed them sorted both alphabetically by DRIT term and numerically by fields/groups. Since they are separate systems of categorization, it should not be surprising that occasionally more than one set of fields/groups (more than one COSATI code) is associated with a single DRIT term.

Software currently in use at NATO is unable to represent 6 digit subject codes. For this reason, a scheme to represent COSATI codes in 4 alphanumeric codes has been specified. The field is left intact in the first two numeric digits. The group is then encoded into a single upper case alphabetic character using A to Z to represent field values of 1 through 26. The subgroup is then encoded similarly. The COSATI code given above of 11/05 for Textiles is thus encoded as 11E. This allows the user to select on the field with some instruction, and all parts of the code with some more instruction and (usually) a reference card. This workaround is designed strictly to get around a limitation in vendor software.

2.7 Related Terms

DRIT does not have related terms as such. Scope notes sometimes (78 times for 13430 posting terms) indicate a "see" or "see also". In 3 cases the scope note indicates a "use" which may also be construed as a related term. DRIT did have related terms at one time (the 1975 edition), but they were later removed because they were not deemed to be chosen and structured in a rigorous enough manner. For this reason, it has been deemed desirable to interate NASA related terms into DRIT for generation of a NATO thesaurus. For more information, see below (sections 3 and 5).

SECTION 3: NASA THESAURUS

3 DRIT + NASA = NATO

NATO's interest in the NASA thesaurus stems primarily from its related terminology: related terms as well as broader and narrower terms. How NASA related terms might be integrated into the DRIT for use by NATO is considered below (section 5). This section will build toward that subject by considering the structure of the NASA thesaurus.

3.1 Lexical format of NASA descriptors

The NASA thesaurus (Reference 1.2.c) is a technical thesaurus broadly similar to DRIT. Like DRIT it focuses on posting terms, only they're called "postable". [The difference in meta-terminology suggests the need for a "thesaurus thesaurus"; as it happens, there is such a document (Reference 1.2.e) but its suggested usage is not yet universally followed.] Four examples of NASA terms are A-3 AIRCRAFT, ATOMIC ENERGY, HALLEY'S COMET, and HUMAN FACTORS ENGINEERING.

As with DRIT, terms are single or multiple words and case distinctions are not significant. The character set is somewhat wider than DRIT's. Characters currently in use are the monocase alphabetic characters A through Z, the parentheses "(" and ")", the dash "-", the apostrophe "'", the slash "/", the ampersand "&", the period ".", and the space used to separate words. [It's probably also true that terms cannot begin with a numeric digit; none currently do.] Four more examples of NASA terms are AIRBORNE/SPACEBORNE COMPUTERS, PAYLOAD DEPLOYMENT & RETRIEVAL SYSTEM, U.S.S.R. SPACE PROGRAM, and VAX-11/780 COMPUTER.

NASA terms are limited to 42 characters. When the term is longer it is abbreviated and/or truncated, then spelled out in full in the scope note. For example ATMOSPHERIC & OCEANOGRAPHIC INFORM SYS is spelled out in the scope note as ATMOSPHERIC & OCEANOGRAPHIC INFORMATION SYSTEMS. The median length is 29 characters. There are about twice as many multiple word terms as there are single word terms.

3.2 Integrity constraints of DRIT descriptors

Like DRIT terms, NASA terms are unique and require no inherent ordering. Therefore they too form a mathematical set. This will be found useful below (section 5) when integrating NASA related terminology with DRIT terms; the two "universes" will be modelled and manipulated using the tools and properties of set theory.

3.3 Relationships between descriptors

3.3.1 Posting Terms and Use References

Like DRIT, NASA specifies a set of use references and reciprocal used for listings. However, NASA does not have a use combination or its reciprocal used for combination.

3.3.2 Broader and Narrower Terms

Like DRIT, NASA establishes reciprocal broader/narrower relationships between pairs of posting terms. Like DRIT, NASA allows both many narrower terms for a term and many broader terms for a term. Therefore, hierarchy and polyhierarchy are part of NASA too. NASA hierarchies tend to be broader than DRIT but not as deep; the degree to and consistency with which this is true has not been ascertained, however.

3.4 Scope notes

Like DRIT, NASA provides scope notes for its postable terms. Also like DRIT, only about 1 in 20 postable terms have a scope note at the present time.

3.5 Subject codes

NASA provides numeric subject codes similar to DRIT. They are four-digit codes organized in a manner similar to DTIC's fields and groups. There are a total of 34 fields.

3.6 Related Terms

NASA does provide a wealth of related terms. As with broader/narrower terms, related terms are all postable terms. This alone makes them very different from DRIT use references, which are never posting terms. All NASA related term relationships are reciprocal; if A is a related term to B, then B is a related term to A. This means that for the 17250 postable terms, the 52732 relationships actually provide for 105464 basic mappings. But this is only the start; NASA postable terms are deeply interrelated. This is diagrammed in Figure 9.

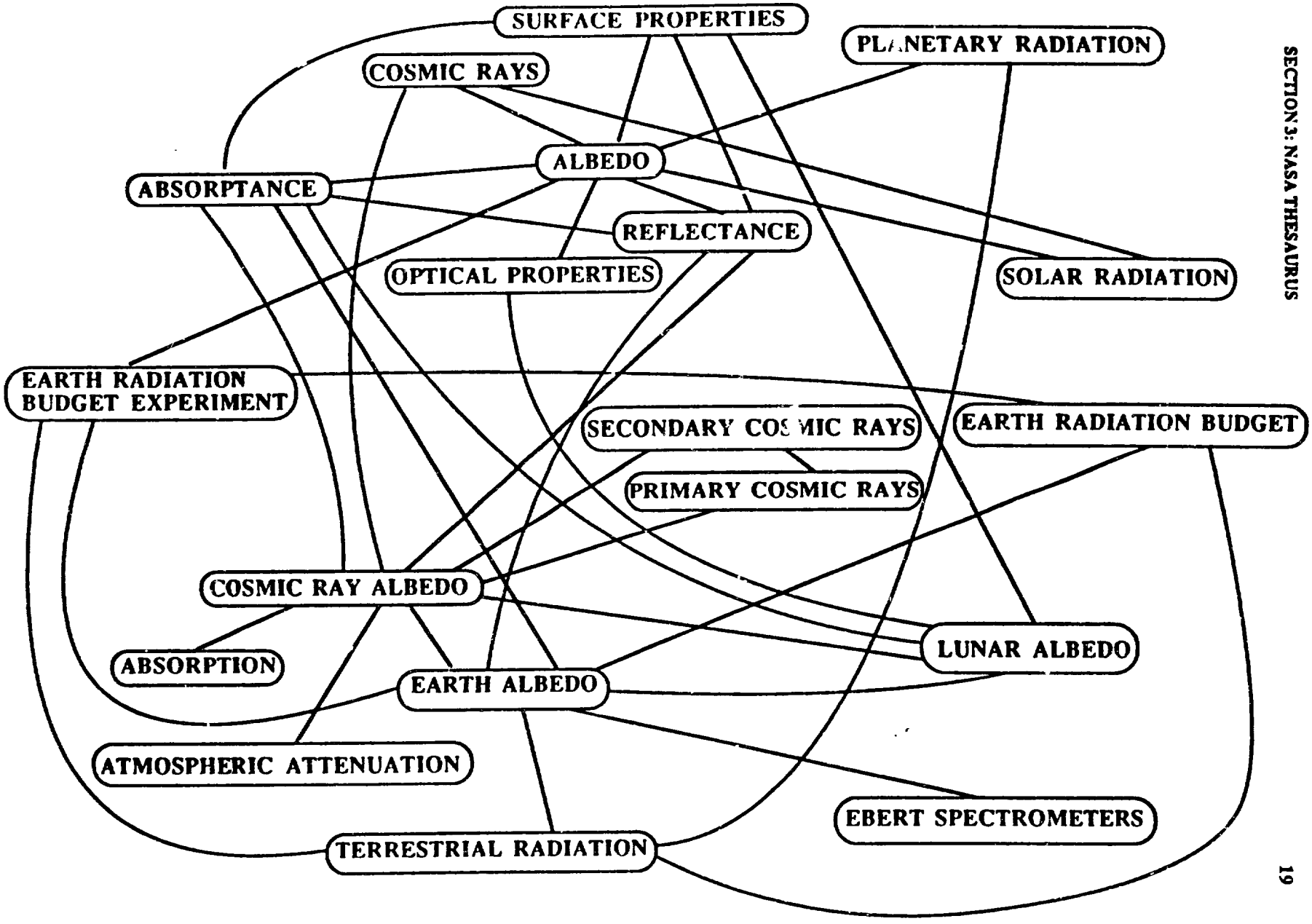


Figure 9. NASA related terms

Figure 9 is a very small piece of NASA related terminology. It displays only the relationships between the terms shown. Had it also attempted to show the relationships between those terms and the other terms in the thesaurus, the diagram might well be illegible. Since related terms, by definition, are neither broader nor narrower, there is no interpretation attached to the top versus bottom placement of terms on the page.

Figure 9 demonstrates that NASA related terms generate a network of relationships that quickly becomes very complex. An important aspect is that each term's related terms form a set, and the intersection of any two of these sets is usually non-empty. For instance, ALBEDO is a related term for both COSMIC RAYS and SOLAR RADIATION. The intersection of the set of the related terms for COSMIC RAYS and the set of related terms for SOLAR RADIATION is the term ALBEDO. This will become important when considering how to integrate NASA related terms.

SECTION 4: ASSIGNMENT OF COSATI CODES TO THE DRIT

4.1 General

As noted above (section 2.6), DRIT has modified COSATI codes for use to the point where it's more appropriate to call them just "fields and groups". However, the use of the term "COSATI codes" is likely to continue if for no other reason than convenience. Therefore in this section the two terms will be used interchangeably.

4.2 Compiling up-to-date information

Since the last edition of the subject categorization guide, about 100 new posting terms have been added to the DRIT for its 1989 edition. In addition, some terms have been retired, and some shifts in meaning have occurred. It has therefore become necessary to obtain updated information on assignment of COSATI codes to DRIT terms. This has been provided, and the keying has been accomplished. This process is not error free but has proceeded at a high level of accuracy.

4.3 Providing a validating data entry interface

The keying is facilitated by providing the data entry person with an interface that allows entry with a minimum of keystrokes but validates basic integrities. This has been performed. The interface allows searching and review of assigned codes, online update, and fast keying of new codes. When more than one code is assigned to a single DRIT term, the interface provides an alternate mode which allows the 2nd through nth code to be assigned. All other times (the common case) a faster mode is used that auto-advances to the next term after each code is assigned or updated.

4.4 Reviewing the assigned codes

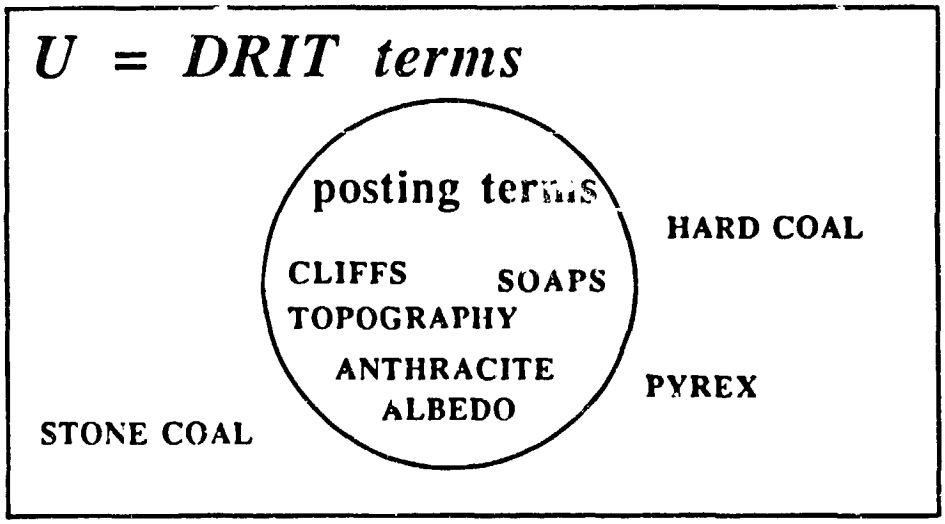
It is finally necessary to review the COSATI codes assigned by DTIC before accepting them as part of the NATO thesaurus. To facilitate this process it is useful to provide a printed listing of the DRIT sorted by COSATI code which shows hierarchies along with their codes. This will permit cross checking both within each hierarchy and within each COSATI code.

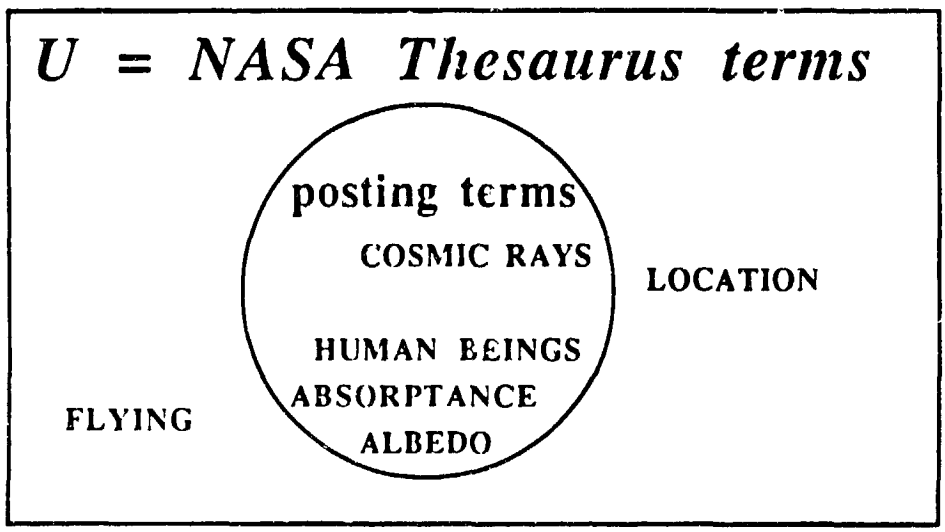
-

SECTION 5: INTEGRATING NASA RELATED TERMS INTO THE DRIT

5.1 The plan: give NASA's related terms to DRIT

The only part of the NASA thesaurus that is to be added is related terms. The means to this end is to connect each DRIT posting term to a NASA postable term. From there, the NASA related terms for that NASA postable term can be added as DRIT related terms for that DRIT posting term. The NASA related terms are NASA postable terms, but they do not become DRIT posting terms. Once again, a Venn diagram provides a more concise description of the task to be performed. A series of three Venn diagrams is provided in Figure 10.





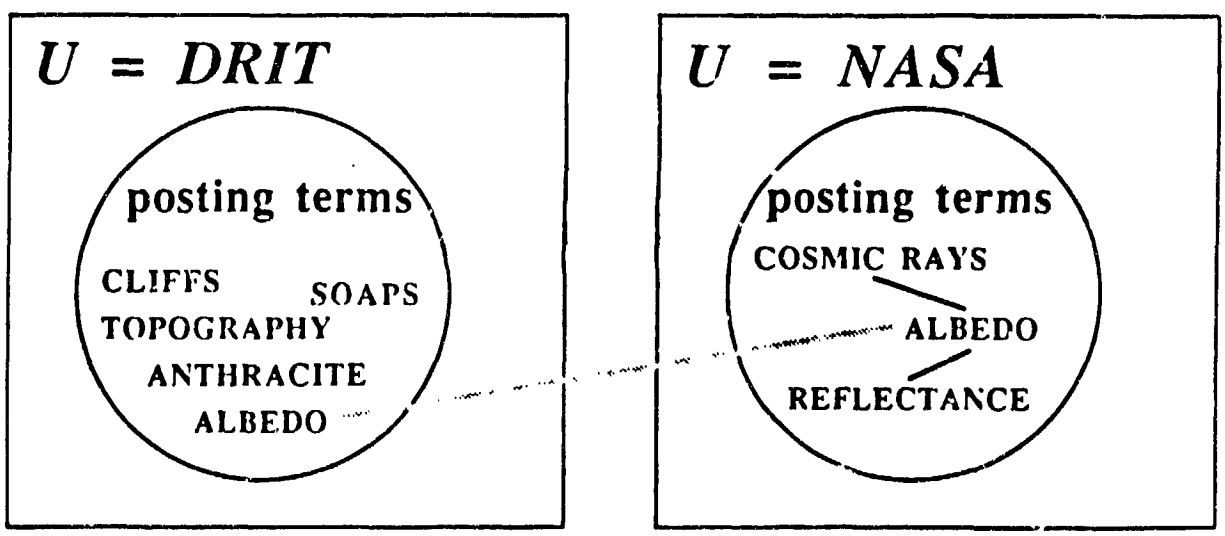


Figure 10. two universes

5.2 The mechanism: subject switching

NASA has long had an interest in integrating thesauri; the NASA STIF maintains a repository of scientific and technical information of interest to the space community. The collection currently contains over three million documents. Of these, a significant percentage come is previously indexed material from DTIC. By integrating the DRIT with NASA thesaurus, the STIF has made it possible to avoid duplication of effort in indexing and abstracting.

The main tool in accomplishing this has been the subject switching file (see Reference 1.2.a). The subject switching file is a table of connections between DRIT and NASA posting terms. This supports table-driven programming that takes DRIT posting terms and finds equivalent NASA postable terms. This would appear at first glance to be exactly what is needed to connect DRIT posting terms to corresponding NASA postable terms. Indeed, it is a reasonable basis. Careful analysis of the development and operational use of the subject switching file reveals three problems to be overcome, however.

5.2.1 Subject switching is not vocabulary matching

The first problem is that subject switching is not vocabulary matching. Subject switching converts the *index set* of terms assigned to a given document from the source vocabulary to the target vocabulary. In theory it is possible that not a single term in the source vocabulary (DRIT) will be converted one-for-one into a corresponding single term in the target vocabulary (NASA). In practice this would be rare, however. The common case is that some terms will be translated one-for-one and some terms will be converted in groups and subgroups. This still opens up the possibility of using the one-for-one table entries in the subject switching file.

5.2.2 The term in an index set is not the term in the thesaurus

More generally, however, translating the set of terms assigned to a given article is the process of translating each source term in the presence or absence of other source terms to one or more target terms. The subject switching file orders entries in the table so that the presence of multiple source terms is considered first. For example, if the DRIT term ABLATION is assigned to the document and the DRIT term NOSE CONES is also assigned, then the NASA term ABLATIVE NOSE CONES will be chosen. If the DRIT term ABLATION is assigned and the DRIT term NOSE CONES is not assigned, then the NASA term ABLATION will be chosen. The solution here has been careful analysis into the consequences of taking the simplest one-for-one translation in all cases. It turns out that this is an acceptable method.

5.2.3 The NASA subject switching file is still using the 1987 DRIT

This was an inconvenience, not a major technical problem. It was compounded, however, by the fact that the machine aided indexing preprocessing strips parentheses as an artifact of its phrase matching preprocessing phase. Of course, s noted above (section 2.1) parentheses are significant in DRIT terms. A fairly large (545 rows) exception table was compiled by hand to match back each DRIT term to its 1989, parenthesized, form.

5.3 Straightforward cases and otherwise

After applying the transforms specified above (section 5.2.3) and the simplifying assumptions described above (section 5.2.2), it turns out that 9911 of the 13430 DRIT posting terms can be matched with a corresponding NASA term. These are all straightforward cases and are handled as such: the set of NASA related terms for that NASA term are now simply assigned to the DRIT term. They do not become posting terms in DRIT or NATO thesaurus. They simply allow users of the NATO thesaurus to find appropriate keywords without having already to know them.

5.3.1 Count of straightforward cases

Of the remaining 3519 cases, about 1000 have no equivalent NASA postable terms, either singly or in combination. This leaves about 2500 cases that might benefit from related terms if there were a rationale to connect via subject switching. Two possible rationales were considered.

5.3.2 One approach for non-straightforward cases

The first approach simply took all possible NASA postable terms that corresponded in any way to the DRIT term, specified all the sets of NASA related terms resulting from each NASA postable term, took the union of all these sets, and returned the terms in this union as related terms to the DRIT term. (Recall that set union by definition eliminates duplicates, so that the high complexity of NASA related terminology doesn't generate errors). If many DRIT terms combined to fewer NASA terms, the related NASA terms for each NASA term were still returned.

This approach was sampled randomly and found wanting; too many related terms found this way are very faintly related to the original DRIT term. An example is the subject switching table entry that says if the DRIT term INTELLIGENCE is assigned and the DRIT term SPACE ENVIRONMENTS is assigned, the NASA term EXTRATERRESTRIAL INTELLIGENCE should be assigned. According to this approach all related terms for EXTRATERRESTRIAL INTELLIGENCE should be assigned back to each DRIT term. One assignment that would be made, therefore, is UNIDENTIFIED FLYING OBJECTS, which is a related term for EXTRATERRESTRIAL INTELLIGENCE, would become a related term for INTELLIGENCE. This does not seem like a good choice.

5.3.3 Another approach for non-straightforward cases

A second approach was also tried. It was the same as the first approach only instead of the union of all sets of resulting related terms, the intersection was performed. Again, by definition this eliminates duplicates. This would eliminate the bad choice found in the previous example. It would also eliminate quite a few good choices. Therefore this approach was also found wanting.

Therefore, at this point a hybrid approach is being used. The related terms currently being submitted on machine-readable media (9-track magnetic tape) are only those found via straightforward means. A minor modification is made if the NASA postable term is a variant of the DRIT term (e.g. spelled differently); in that case it is also listed as a related term. The related terms being submitted for human inspection on

paper media are those found via non-straightforward means, but the NASA terms used in translation are identified next to each additional unique related term found via that term.

28

SECTION 6: TRANSLATING THE DRIT INTO FRENCH

6.1 General information

This part of the project is in its very preliminary stages. The status right now is that agreement has been arrived at as to the format and layout of a machine-readable translation of the DRIT posting terms into French. It will then be possible for searches in either language to provide results in either language. Note that this does not translate retrieved citations for the user; it merely reduces the size and scope of the potential translation burden. This is still a highly desirable goal, since it would still to some degree facilitate international cooperation in scientific and technical interchange.

6.2 Translation method

The details of the translation method have not been finalized. It is anticipated, however, that automated translation software will perform a "first pass", and then human checkers will verify and correct the output. To facilitate the human effort, a special listing of the DRIT has been prepared the provides both the "thesaurus format" listings and also embedded hierarchies; that is, hierarchy listings of each term next to its alphabetically sorted listing. This has already proven valuable in getting good translations. Translation staff at NATO have identified an example of SUBMARINE NOISES. This term was taken too literally by the automated translation software, to mean any noise made under water, such as whale noises. In its hierarchies of ACOUSTICS and NOISE, it becomes clear that the term means noises made by submarine vessels. This will be based on the understanding gained of DRIT in work performed so far. The information provided above (section 2) will provide a framework for specifying the relationships between French and English terms and broader/narrower, deprecated/preferred, and related terms.

SECTION 7: PREDECESSOR DOCUMENTS TO THIS REPORT

7.1 Predecessor Activity

The following reports document activity directly related to the NATO Thesaurus Project.

Reference (1), 1989, describes the approach to the development of a common thesaurus for use by national and international Defense Scientific and Technical Information (STI) organizations to facilitate the exchange of information.

Reference (2), 1988, is a discussion of the ramifications of technology applications for a global network that encompasses scientific and technical information across the globe.

Reference (3), 1987, describes the development and implementation of an integrated, functional scientific and technical information network to access national and international information sources, covering technologies for networking, accessing, interfacing, and processing information aggregated from those diverse sources.

7.2 Documents

(1) **Terminology Strategies for International Information Exchange.**
Gladys A. Cotter, Defense Applied Information Technology Center, Alexandria, Virginia; and
Walter R. Blados, Deputy for Scientific and Technical Information, Secretary of the Air Force, The Pentagon, Washington, DC.
DAITC/TR-89/9, August 1989, AD-214 147.

(2) *Global Scientific and Technical Information Network.*
Gladys A. Cotter.
Online Information 88, 12th International Online Information Meeting, 6-8 December 1988, London;
pp. 611-618.

(3) *Information Retrieval Systems Evolve - Advances for Easier and More Sucessful Use.*
Gladys A. Cotter.
Paper 5 in **AGARD Conference; Barriers to Information Transfer and Approaches Toward Their Reduction.**
AGARD-CPP-430, 1987.