

CONFERENCE OF EUROPEAN STATISTICIANS

UN/ECE Work Session on Statistical Metadata
(Washington, D.C., United States, 28-30 November 2000)

Topic (i): Statistical metadata for dissemination

A WAY WITH WORDS: THESAURI ASSISTED SEARCHING

Submitted by University of Essex, United Kingdom¹

Contributed paper

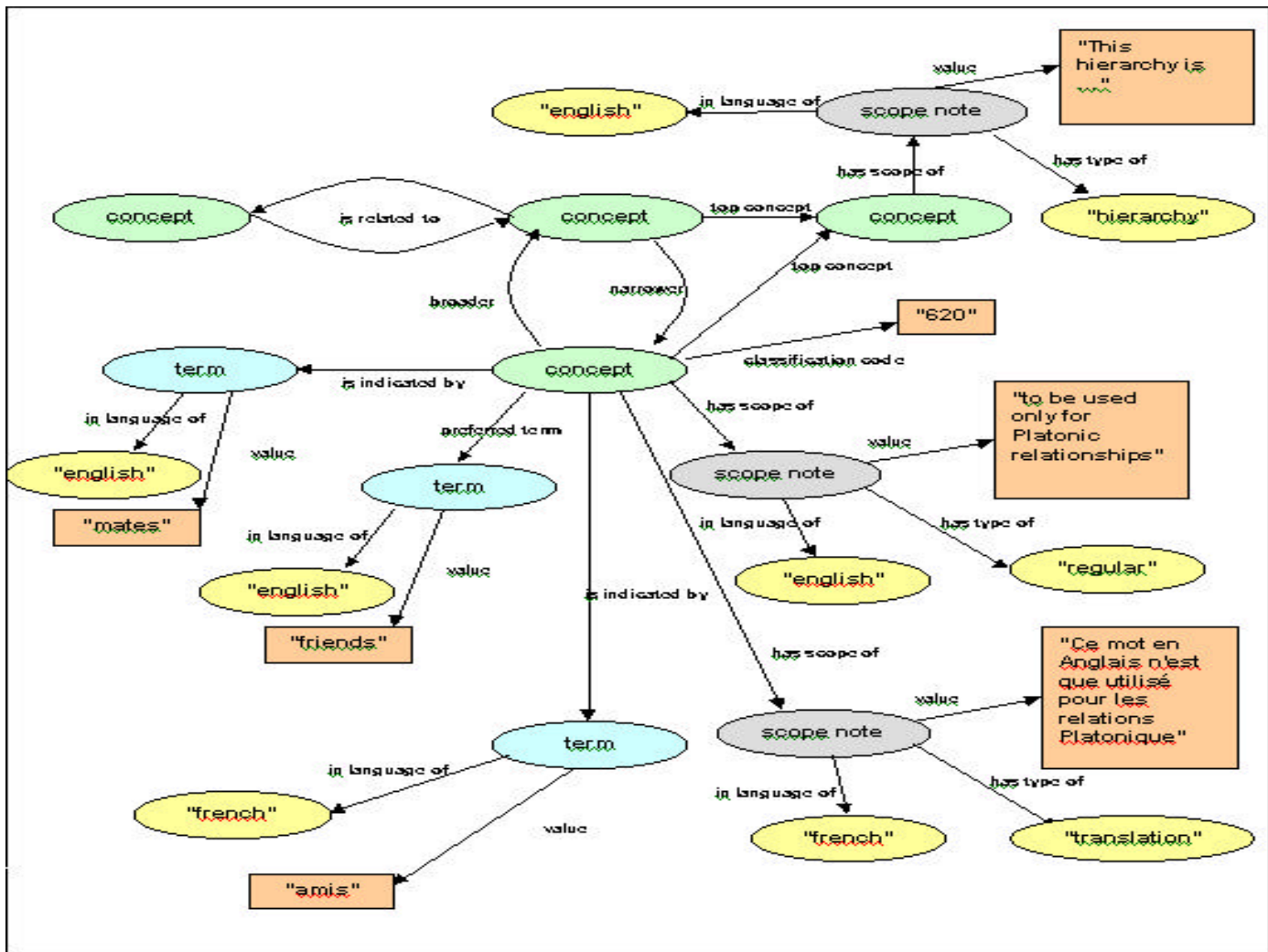
I. INTRODUCTION

1. This paper will discuss the use of thesauri as an aid to searching for on-line resources. In particular it will draw on the example of the Humanities and Social Science Electronic Thesaurus (HASSET) and its present use in the UK Data Archive's search engine BIRON (Bibliographic Information Retrieval ON-line) and its future use in the European project LIMBER (Language Independent Metadata Browsing of European Resources).

2. My introduction to thesauri, as with many others, was through my obsession with crosswords where Roget's thesaurus proved invaluable. Consider the clue for a six-letter word beginning with "L":-pliant carriage for the artillery. Here the Roget's entry for "pliant" appears under the heading "softness" which also lists two possible solutions "lissom" and "limber". The second of these also appears in the "vehicle" listing. However, there is no indication from its immediate neighbours "tram" and "tumbrel" that a limber is "a two-wheeled vehicle to which a gun or caisson may be attached", the definition from Webster's dictionary.

3. Although a useful tool, Roget's thesaurus is not the type of thesauri to be discussed here. Modern hierarchical thesauri are much better at organising knowledge through their ability to connect one concept to another in various ways through synonyms, narrower, broader and related links. The resulting tool allows for consistency of indexing, browsing of subject concept trees, automatic substitution of terms in free-text searches, automatic narrowing or broadening of searches, ranking of results, high relevance, cross-domain and multi-lingual searches.

¹ Prepared by Ken Miller.



II. THESAURUS ELEMENTS

4. The following section describes the building blocks on which all the functions listed above can be built.

5. Preferred Term:- Every concept in a thesaurus is assigned a term, consisting of a single word or phrase, which is considered the most apt to describe the concept. The term is unique within the thesaurus and qualifiers can be used to ensure this, e.g. LABOUR (BIRTH) and LABOUR (WORK). It is the preferred term that is always applied when indexing thus ensuring consistency and high relevance when searching.

6. Use For (UF):- Every preferred term can have "use for" relationships with synonyms describing the same concept. For example EMPLOYMENT uf WORK. These act as lead in terms to the preferred term when searching on the controlled vocabulary of the thesaurus or as alternatives in a free text search. They can also be used for equivalent terms in different languages or different domain thesauri.

7. Scope Note (SN):- Although the display of a thesaurus entry within the subject concept tree is usually sufficient to convey meaning, sometimes ambiguity can arise. In these cases a scope note can be attached to a preferred term. For example HOMEWORKERS sn WORKERS WORKING IN THEIR

OWN HOMES ON MATERIALS SUPPLIED BY THE EMPLOYER. This obviously aids assignment of indexing terms and selection for searching.

8. Notation Code (NC):- In some thesauri a notation can be assigned to each preferred term, so that the thesauri can act as a subject classification useful for physical location of resources. For Example ECONOMIC AID nc L84.50.20

9. Narrower Term (NT):- Every preferred term can have "narrower term" relationships to other preferred terms. These narrower terms have concepts narrower in scope to the concept covered by the preferred term being assigned the relationship. For example TRADE nt DOMESTIC TRADE. Hence when browsing the thesaurus a search can easily be switched to a narrower concept that more precisely covers the subject area required. Automatic inclusion of a search terms narrower terms is also a way of widening a search.

10. Broader Term (BT):- Every preferred term can have "broader term" relationships to other preferred terms. These broader terms have concepts broader in scope to the concept covered by the preferred term being assigned the relationship. For example TRADE bt ECONOMICS. This type of relationship is especially useful when browsing the thesaurus.

11. Related Term (RT):- Every preferred term can have "related term" relationships to other preferred terms. A related term relationship is used to make associative links between conceptually related terms within different hierarchies of the thesaurus. Related terms are those which have a close conceptual link to the preferred term but which are not in an equivalence relationship (UF) or in a hierarchical relationship (BT/NT). For example TRADE rt CONSUMPTION.

12. Top Term (TT):- Some thesauri also give every preferred term relationships to the very broadest concept of the hierarchies in which term appears, e.g. DOMESTIC TRADE tt ECONOMICS. These "top terms" are obviously terms that have no "broader term" relationships to any other concept.

III. THESAURUS DISPLAYS

13. These relationships above can be displayed in various ways, the alphabetic listing, below, show all the relationships to a single preferred term.

TRADE NC=N65/75

SN TO INDEXERS: USE A MORE SPECIFIC TERM IF AVAILABLE
 UF COMMERCE
 NT DOMESTIC TRADE
 INTERNATIONAL TRADE
 MARKETING
 RETAIL TRADE
 TRADING HOURS
 WHOLESALE TRADE
 BT ECONOMICS
 TT ECONOMICS
 RT COMMERCIAL LAW
 CONSUMPTION
 MARKETS (ECONOMICS)
 PRICES
 SUPPLY AND DEMAND

14. The hierarchical display reveals whether the listed narrower terms have further narrower concepts and whether there are additional broader terms in between those shown and the top term. It does not display any related terms.

ECONOMICS
 TRADE
 DOMESTIC TRADE
 INTERNATIONAL TRADE
 ECONOMIC AGREEMENTS
 EXPORTS/IMPORTS
 FREE TRADE
 INTERNATIONAL COMPETITION
 PROTECTIONISM
 MARKETING
 AGRICULTURAL MARKETING
 SELLING
 INERTIA SELLING
 SALE OF PERSONAL POSSESSIONS
 VENDING MACHINES
 RETAIL TRADE
 TRADING HOURS
 WHOLESALE TRADE

15. The classified listing tries to include the related terms along with the narrower/broader relationships in a subject related hierarchy.

N65/75 TRADE **
 N65 COMMODITIES
 N65.05 TRADE ASSOCIATIONS
 N66 SUPPLY AND DEMAND
 N66.10 SUPPLY
 N66.40 DEMAND
 N66.40.01 BUSINESS ORDERS
 N67 MARKETS (ECONOMICS)
 N67.40/60 MARKET STRUCTURE
 N67.50 MONOPOLIES
 N67.60 ECONOMIC COMPETITION
 N68 CONSUMPTION
 N68.10 CONSUMER GOODS
 N68.10.50 SECOND-HAND GOODS
 N68.20 PURCHASING
 N68.20.10 SHOPPING
 N68.20.10A SUNDAY SHOPPING
 N68.50 SOCIOLOGY OF CONSUMPTION
 N69 PRICES
 N69.05 TARIFFS
 N69.10 PRICE POLICY
 N69.20 PRICE CONTROL
 N69.30 ECONOMIC VALUE
 N69.30.20 DEPRECIATION
 N69.60 INFLATION
 N69.60.10 DEFLATION

16. An additional display, sometimes available in thesauri, is a keyword in context (KWIC) listing of all preferred terms and synonyms that contain an entered word or words from an existing preferred term or synonym. For example a KWIC listing generated from the preferred term PRICE CONTROL could display the following:-

CONSUMER PRICE INDEX
 PRICE CONTROL
 PRICE POLICY
 PRODUCER PRICE INDEX
 RETAIL PRICE INDEX
 ARMS CONTROL
 BIOLOGICAL CONTROL
 BIRTH CONTROL
 BUDGETARY CONTROL
 CONTROL OF POLICE
 CONTROL OF THE ENVIRONMENT
 DRUG CONTROL

Use in search engines

17. The relationships and the displays within a thesaurus have been described above, so how now can we start to use them in earnest to retrieve information.

18. Browsing:- The simplest use of thesauri in search engines is as a basic browsing facility. Here either the top terms of each hierarchy in the thesaurus are displayed as a starting point or a simple search box returns a KWIC listing of terms generated from the words in the entered string. The alphabetic or hierarchical listings are usually displayed for the user just to select individual terms to search on. This can be as basic as a cutting and pasting operation to an external search engine or pre-set searches behind each thesaurus term. More complex features can include a shopping basket option to build up searches on two or more selected terms, either with Boolean "and" or "or" between each or the ability to form more complex Boolean expressions before passing on to the actual search engine.

19. Free-Text Search:- Simply using a preferred term and all of its synonyms in a free text search is one way of using a thesaurus but usually results in a large amount of irrelevant retrievals. This is due to the fact that the semantics of the terms retrieved are not defined. Hence a search on labour + birth could result in a resource describing "the birth of the labour party" even if proximity factors are used as well. A more refined use of thesauri involves the weighting of the relationship terms. For example the term itself and all synonyms would be weighted 10, all narrower terms and their synonyms weighted 8, all broader terms and synonyms 6, all related 4 and the top term 2. Searching on so many terms obviously adds to the noise but the weighting allows ranking of results with very high relevance in the top scoring hits. This is similar to the method employed by the CIA to monitor email traffic. Here however each relationship is weighted separately from the main concept of "subversive activity" and actual emails read by CIA staff is under 1% of total Internet traffic.

20. Controlled Vocabulary Search:- It is in this kind of search that the power of a thesaurus can be fully exploited. The fact that the resource itself has been consistently indexed by terms taken from the same controlled vocabulary that is used to search by, ensures high relevance and high retrieval. Since all relationship terms have been applied in a similar consistent manner, means that automatic inclusion of relationship terms to widen searches also results in high relevance and high retrieval. Ranking of hit lists can be as simple as a count of the number of search terms within each resource, although weighting of relationships, as described above for free text searches, gives far better results. A resource that matches a user's requirement can also be used as a basis for finding similar resources, by searching on all the keywords assigned to the selected resource. Where problems can arise is in Boolean "and" searches, especially if the user is trying to locate a narrower concept than reflected in the thesaurus. For example, a user trying to locate a survey in which questions were asked about the salaries of farm workers might

search on a combination of terms from the thesaurus, "wages" and "agricultural workers". If the resource was only indexed at survey level then there is no guarantee, even if both keywords had been applied, that they were used in combination to describe agricultural workers wages. This can be overcome by assigning some link between terms used in combination or assigning keywords at a more specific level.

21. Cross-Domain Search:- This involves the mapping between domain specific thesauri. In the example above the terms "wages" and "agricultural workers" came from a social science thesaurus used to index surveys held in social science data archives. The user however may wish to widen their search across resources held at agricultural archives indexed by another thesaurus but still using the terms from their familiar domain specific thesaurus. Mappings can be made between the two thesauri, such as between equivalent terms "agricultural workers" and "farm workers", "wages" and "salaries" or between combinations of terms so that "agricultural workers" and "wages" is mapped as equivalent to "farm workers salaries". If one thesaurus has more specific concepts narrower to the terms on which equivalence was made, then these can be automatically pulled in for a wider search.

22. Multi-Lingual Search:- This too can involve mapping between domain specific thesauri in different languages, although a single multi-lingual thesaurus does have some advantages. Obviously the mappings have been done already and are usually of stronger equivalence. Also if a notation code exists, searching can be performed on this across languages rather than having to translate the search query to every language. Resources discovered in the user's non-native language, can then have the keywords assigned accurately translated to aid determining the relevance of these foreign resources.

IV. USE OF HASSET IN BIRON

23. The use of the Humanities and Social Science Electronic Thesaurus (HASSET) in the UK Data Archive's search engine BIRON (Bibliographic Information Retrieval On-line) is basically as described under c) above, Controlled Vocabulary Search. Automatic substitution of synonyms and generation of KWIC lists of terms, direct the user to the preferred term of the controlled vocabulary. Further thesaurus-aided searches are available on request. Selection of one or more terms from both the alphabetic or hierarchical displays is available as well as mass inclusion of all related terms and narrower or broader terms to any level.

BIRON 4.21 . *internal*

Using *Agricultural Workers* as preferred term for *Farm Workers*

Subject thesaurus entry: **Agricultural Workers**

Select one or more terms to extend search

The screenshot shows a web interface for the BIRON 4.21 internal thesaurus. At the top, it displays the subject thesaurus entry: **AGRICULTURAL WORKERS**. Below this, a list of related terms is shown in a scrollable box:

- uf.. FARM WORKERS
- uf.. PEASANTRY
- nt.. FEUDAL AGRICULTURAL WORKERS
- bt.. WORKERS
- tt.. ECONOMICS
- tt.. LABOUR AND EMPLOYMENT
- tt.. RESOURCES

To the right of the list is a 'Key to term types' link. Below the list are three buttons: 'New thesaurus entry', 'List KWIC terms', and 'List full hierarchy', each with a question mark icon. To the right of these buttons is a dropdown menu labeled 'from top term' with 'ECONOMICS' selected. Below the buttons and dropdown are two rows of search options:

- Row 1: Include narrower level terms, Include broader level terms, and a dropdown menu for 'extended to search level' set to '1'.
- Row 2: Include all related terms, Search selected term(s), and a 'Reset' button.

24. The terms from any multiple selection are always combined by a Boolean "or" in the proceeding search. The search however can be refined by combining via Boolean "and", "or" or "not" with further searches on one or more terms. The fact that the thesaurus has been developed by the UKDA itself means that the subject coverage fully matches that of the holdings. So if a subject area is not in the thesaurus it means none of the surveys held at the UK Data Archive have data covering that subject. Unfortunately although the index terms from HASSET assigned to the metadata records in BIRON reflect the actual questions asked and the variables contained in the data they are only assigned at survey level. Hence the problems using Boolean "and" outline above in c) can occur with searches. Furthermore there is no ability to rank retrievals or find similar data via the assigned keywords.

V. USE OF ELSST IN LIMBER

25. The aim of the LIMBER project is to provide plug-in tools for search engines to add the full power of multi-lingual hierarchical thesauri. Basically to provide everything outlined in a) through e) above. LIMBER (Language Independent Metadata Browsing of European Resources) is an Information Societies Technology, European Union funded project which seeks to address the problems of linguistic and discipline boundaries which, within a more integrated European environment, are becoming increasingly important. Decision-makers, researchers and journalists need to be provided with a broader, comparative picture of society across the continent, with the social science information often required to be correlated with information from domains such as environmental science, geography and health. This cross-discipline interoperability will be provided via a uniform metadata description. In addition the provision of multilingual user interfaces and the controlled vocabulary of a multi-lingual thesaurus will make these data globally accessible in a range of end user natural languages.

26. The first prototypes will supplement the NESSTAR system (Networked European Social Science Tools and Resources) which uses the DDI (Data Documentation Initiative) metadata standard for social science codebooks and NSDStat statistical engine to display and browse comprehensive metadata records and allow simple analysis of the underlying data.

27. The LIMBER tools will provide browsing of the multi-lingual thesaurus (ELSST) in the user's natural language; thesaurus aided cross-domain, multi-lingual free-text and controlled vocabulary searching with relevance feedback provided in the user's own language. Hit lists will be ranked depending on number of occurrences and position within the metadata record. To aid the indexing of the metadata a semi-automatic indexing tool will be developed to assign keywords at various places throughout the metadata. Although initially concentrating on the DDI and ELSST the architecture and standards adopted will allow for mappings between metadata standards and domain specific thesauri hence ensuring a completely resource, language and thesaurus independent system.

VI. CONCLUSIONS: IT'S NOT WHAT YOU SEARCH IT'S THE WAY THAT YOU SEARCH IT - AND THAT'S WHAT GETS RESULTS!

28. Up until recently this was certainly the philosophy behind web search engines, and you did get results, lots of results. The emergence of the Dublin Core as a metadata standard for web resources and interoperability has led to the widely recognised need for such a core element set. However, few applications have found that the 15 elements satisfy all their needs and most applications require further mechanisms to refine or qualify metadata elements or their values. This has led to the emergence of several domain specific metadata initiatives and the adoption of domain specific controlled vocabularies in several languages. However, this leads to interoperability problems between the metadata standards, the controlled vocabularies and the languages, which are major obstacles to resource discovery.

29. The adoption of multilingual thesauri and mappings between metadata standards and the controlled vocabularies can overcome these obstacles. The dream of being able to use a simple Alta-vista style search but having a relevant hit-list sensibly ranked, with indication in your own language as to why the resources were selected might just become a reality.