Meaningful Relationships

by Ken Miller [*]

INTRODUCTION

The Data Archive's thesaurus, HASSET (Humanities and Social Science Electronic Thesaurus), is based upon the UNESCO thesaurus compiled by Jean Aitchison (Paris; UNESCO 1977) and has been built up over 18 years so that its coverage reflects the subject matter of the 5,000 datasets held at The Data Archive.

This paper will describe the construction, maintenance and use of the thesaurus as a controlled vocabulary for indexing and a retrieval tool in The Data Archive's on-line catalogue BIRON (Bibliographic Information Retrieval ON-line),. It will also outline The Data Archive's proposed developments for HASSET as an on-line thesaural resource for the social science community in general and as a multilingual free-text retrieval tool within, among others, the NESSTAR (Networked European Social Science Tools and Resources) project.

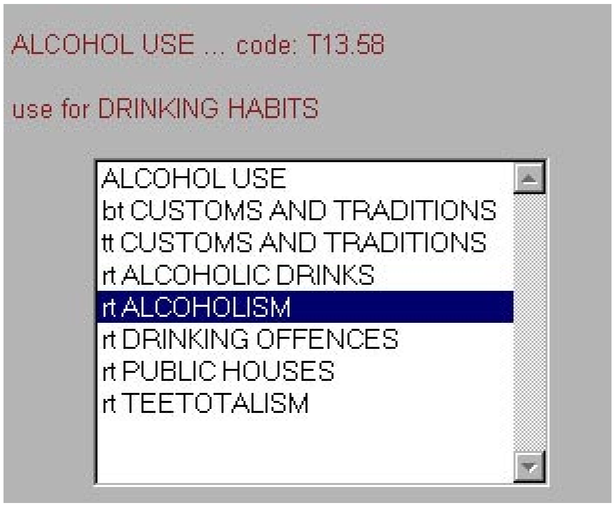
THESAURI

Dictionary definitions of thesauri describe them as "a storehouse of information" e.g. a dictionary, or "a list of concepts or words arranged according to sense" e.g. Roget's, or "a list of concepts or words chosen for use in indexing" e.g. the UNESCO thesaurus.

HASSET is all this and more and is why we at The Data Archive consider it just that, a Huge ASSET. Its use as a controlled vocabulary means that every dataset whose question or variable covers the same subject material will be indexed by the same concept term. The structured relationships allow the indexer to view candidate terms within a concept hierarchy. The fact that it is machine readable allows instant, easy and consistent maintenance and flexibility in displaying terms in various different ways. It also means that it can be used as a retrieval tool in BIRON helping the searcher to better define, expand or focus their search

HASSET

There are six basic relationships between the terms held in the thesaurus and these are held in one database table with the simple format of CONCEPT TERM - relationship type - RELATIONSHIP TERM. They are 1) Use 2) Use For - UF 3) Narrower Term - NT 4) Broader Term - BT 5) Top Term - TT 6) Related Term - RT.



ALCOHOL USE ... code: T13.58

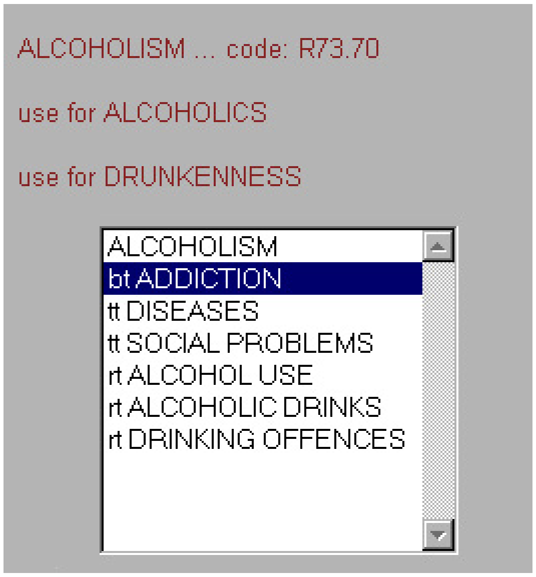
use for DRINKING HABITS

ALCOHOL USE
bt CUSTOMS AND TRADITIONS
tt CUSTOMS AND TRADITIONS
rt ALCOHOLIC DRINKS
rt ALCOHOLISM
rt DRINKING OFFENCES
rt PUBLIC HOUSES
rt TEETOTALISM

Hence :-

DRINKING HABITS - use - ALCOHOL USE
i.e. “drinking habits” is a non-preferred synonym of the preferred term “alcohol use”
There will also be the reciprocal entry :- ALCOHOL USE - use for - DRINKING HABITS

ALCOHOLISM - broader term - ADDICTION
with the reciprocal entry :- ADDICTION - narrower term - ALCOHOLISM
i.e. There is a narrower concept “alcoholism” to the subject term “addiction”

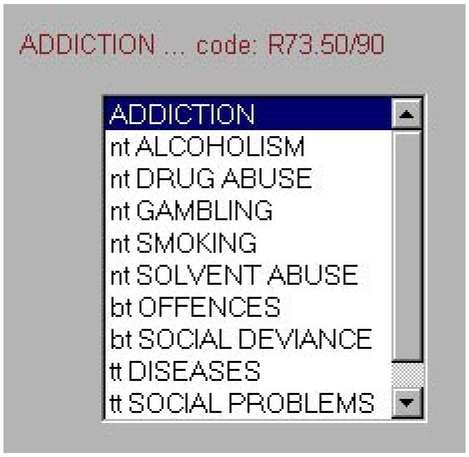


ALCOHOLISM ... code: R73.70

use for ALCOHOLICS

use for DRUNKENNESS

ALCOHOLISM
bt ADDICTION
tt DISEASES
tt SOCIAL PROBLEMS
rt ALCOHOL USE
rt ALCOHOLIC DRINKS
rt DRINKING OFFENCES



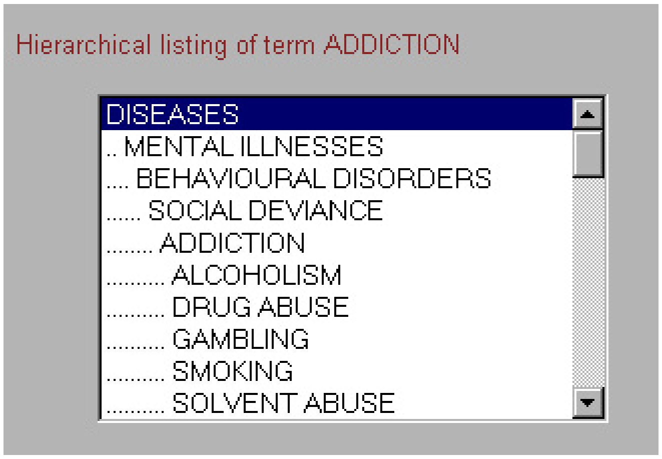
ADDICTION ... code: R73.50/90

ADDICTION
nt ALCOHOLISM
nt DRUG ABUSE
nt GAMBLING
nt SMOKING
nt SOLVENT ABUSE
bt OFFENCES
bt SOCIAL DEVIANCE
tt DISEASES
tt SOCIAL PROBLEMS

Both subject terms “alcoholism” and “addiction” are in two hierarchies, one from the top term “diseases” and one from the top term “social problems”. Hence the following entries are found in the database table :-

ALCOHOLISM - top term DISEASES	ADDICTION - top term - DISEASES
ALCOHOLISM - top term SOCIAL PROBLEMS	ADDICTION - top term - SOCIAL PROBLEMS

N.B. there is no reciprocal entry for a top term relationship, which acts as an aid to understand the scope and meaning of the subject concept under review, and in programming to build up the correct hierarchies.



Hierarchical listing of term ADDICTION

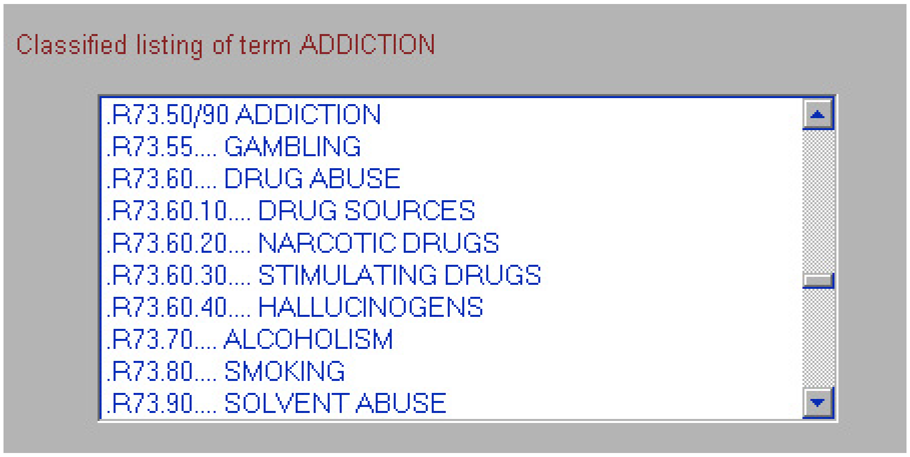
```
DISEASES
.. MENTAL ILLNESSES
.... BEHAVIOURAL DISORDERS
...... SOCIAL DEVIANCE
........ ADDICTION
.......... ALCOHOLISM
.......... DRUG ABUSE
.......... GAMBLING
.......... SMOKING
.......... SOLVENT ABUSE
```

The final relationship is that between two preferred subject terms which are related to each other but are not covered by the NT, BT or TT relationships. The reciprocal entry is also included in the database table. Hence:-

ALCOHOL USE - related term - ALCOHOLISM

ALCOHOLISM - related term - ALCOHOL USE

HASSET does actually have two other database tables, one which holds a textual clarification of the subject term, known as a scope note (SN), and the second holds a classification code which places the subject term in one fixed hierarchy, so that HASSET could be used as a shelving scheme for hard copy documentation, and a marker to show whether the term was taken from the UNESCO thesaurus or is a Data Archive new term.



Classified listing of term ADDICTION

.R73.50/90 ADDICTION
.R73.55.... GAMBLING
.R73.60.... DRUG ABUSE
.R73.60.10.... DRUG SOURCES
.R73.60.20.... NARCOTIC DRUGS
.R73.60.30.... STIMULATING DRUGS
.R73.60.40.... HALLUCINOGENS
.R73.70.... ALCOHOLISM
.R73.80.... SMOKING
.R73.90.... SOLVENT ABUSE

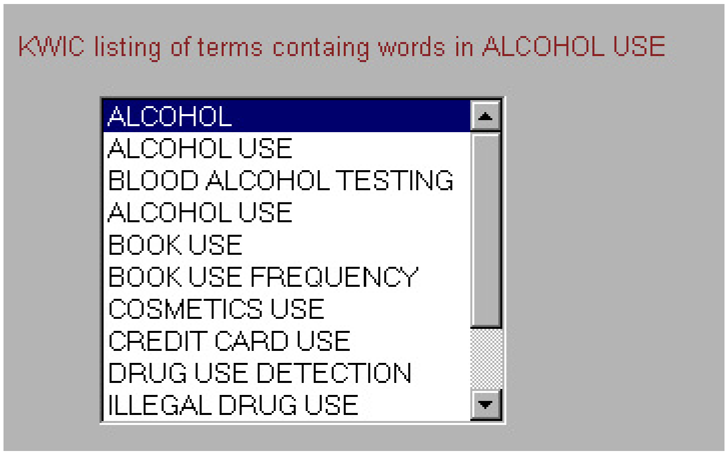
There are, at present, approximately 8,650 terms in HASSET; 2,500 of which are non-preferred terms or synonyms. 38,600 relationships exist between these terms and they form 296 hierarchies. Approximately half of the terms have been taken from the UNESCO thesaurus.

MAINTENANCE & INTERFACES

The tables described above are held in an INGRES database and updates to the thesaurus are performed through 'C' programs, written at The Data Archive, which employ embedded SQL calls to the underlying tables. The same interface also performs the indexing of the actual datasets with terms from the controlled vocabulary.

The program ensures that reciprocal entries are automatically included, terms are correctly positioned in hierarchies with the most appropriate allocation of classification code. The duplication of terms is impossible as is the creation of incorrect relationships between terms.

To aid the allocation of terms to the datasets held at The Data Archive, the indexer has recourse not only to the thesaurus, hierarchical and classification listings described above, but also the scope notes, listings of datasets previously indexed by the term under consideration and a KWIC (keyword in context) listing of words from the candidate term. The example below shows the kwic listing for the term "alcohol use".



KWIC listing of terms containg words in ALCOHOL USE

ALCOHOL
ALCOHOL USE
BLOOD ALCOHOL TESTING
ALCOHOL USE
BOOK USE
BOOK USE FREQUENCY
COSMETICS USE
CREDIT CARD USE
DRUG USE DETECTION
ILLEGAL DRUG USE

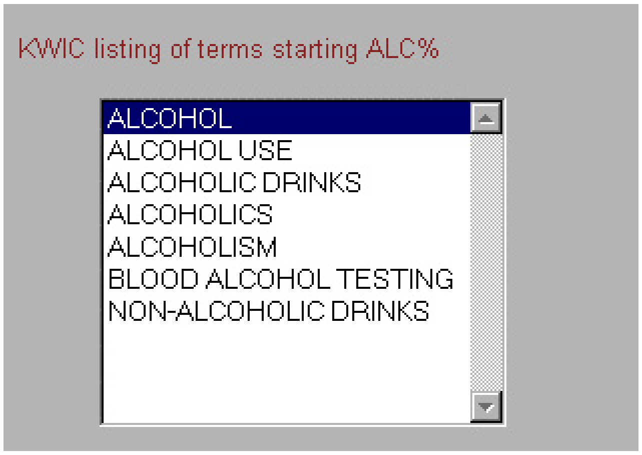
BIRON & HASSET

The WWW interface to both HASSET and BIRON is through dynamically produced html forms from a cgi-bin ‘C’ program with embedded SQL calls to the underlying INGRES database tables.

How then does HASSET aid the searcher of The Data Archive’s on-line catalogue BIRON. First of all the indexing program ensures that the same subject concept in any dataset held is assigned the same controlled vocabulary term. BIRON’s first task then is to point the user to the preferred term, if the keyword searched on is not in itself a preferred term; it does this in three ways.

Firstly it searches the synonyms from the USE and UF relationships to see if it can find a match. If it does it will automatically substitute the preferred term and carry out a search immediately. If it cannot match against a non-preferred term then the program produces a KWIC listing from the word or words in the search term.

Finally, if the second option fails, BIRON will produce another KWIC listing, but this time from progressively truncating the search string until a listing is produced, even if it has to be a list of all terms in the thesaurus. Hence entering a slight misspelling of “alchol” results in :-



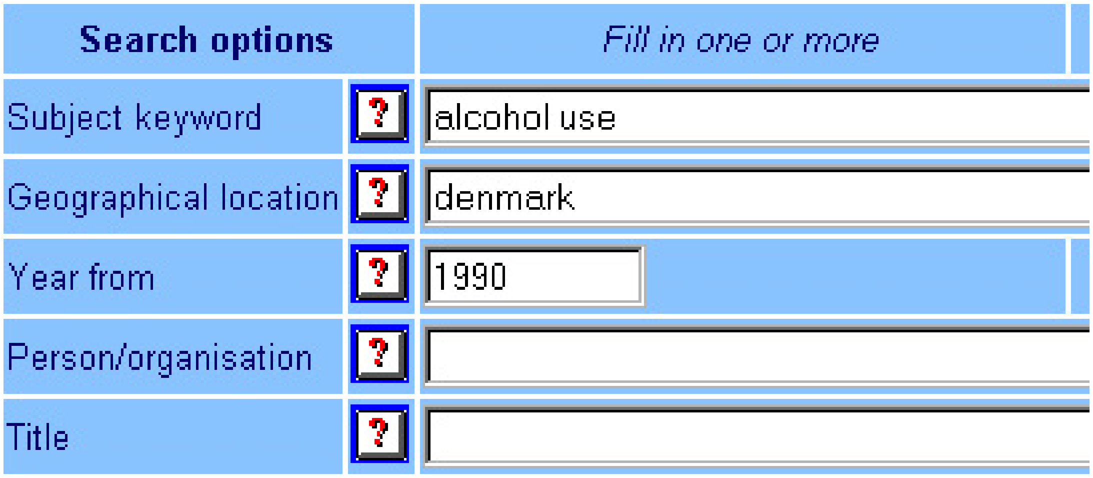
KWIC listing of terms starting ALC%

```
ALCOHOL
ALCOHOL USE
ALCOHOLIC DRINKS
ALCOHOLICS
ALCOHOLISM
BLOOD ALCOHOL TESTING
NON-ALCOHOLIC DRINKS
```

Therefore the searcher is always offered some candidate terms no matter what is entered as the search string. Selection is carried out by just clicking on the required term and then the on “search” icon to perform the search.

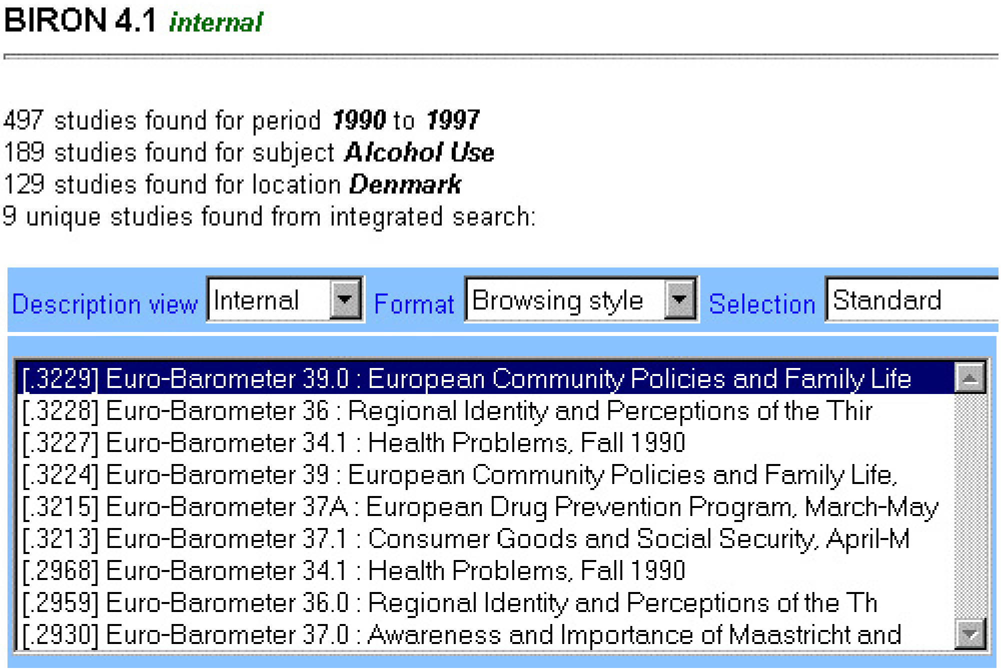
Once a search has been carried out the thesaurus is also available to help the searcher redefine their search through the displays described above, by changing to broader or narrower concepts, adding more terms to their search or combining the results from their present search, so that the datasets retrieved also cover the concept of another subject term or terms displayed.

Consider the following search:-



Search options		Fill in one or more
Subject keyword	?	alcohol use
Geographical location	?	denmark
Year from	?	1990
Person/organisation	?	
Title	?	

Which results in :-



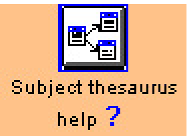
BIRON 4.1 *internal*

497 studies found for period ***1990*** to ***1997***
189 studies found for subject ***Alcohol Use***
129 studies found for location ***Denmark***
9 unique studies found from integrated search:

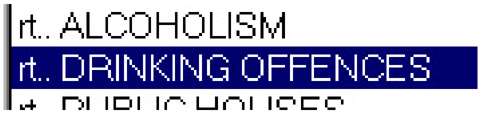
Description view: Internal
Format: Browsing style
Selection: Standard

[.3229] Euro-Barometer 39.0 : European Community Policies and Family Life
[.3228] Euro-Barometer 36 : Regional Identity and Perceptions of the Thir
[.3227] Euro-Barometer 34.1 : Health Problems, Fall 1990
[.3224] Euro-Barometer 39 : European Community Policies and Family Life,
[.3215] Euro-Barometer 37A : European Drug Prevention Program, March-May
[.3213] Euro-Barometer 37.1 : Consumer Goods and Social Security, April-M
[.2968] Euro-Barometer 34.1 : Health Problems, Fall 1990
[.2959] Euro-Barometer 36.0 : Regional Identity and Perceptions of the Th
[.2930] Euro-Barometer 37.0 : Awareness and Importance of Maastricht and

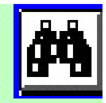
Clicking on the thesaurus help icon displays the thesaural entry for “alcohol use”.

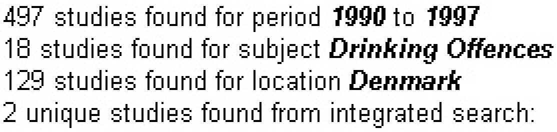


From which you could select the related term



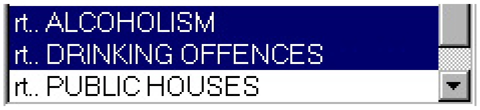
Select the button and click on the icon which results in :-



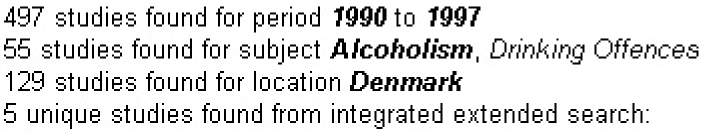


497 studies found for period ***1990*** to ***1997***
18 studies found for subject ***Drinking Offences***
129 studies found for location ***Denmark***
2 unique studies found from integrated search:

Or you could select more than one related term and click on the extend icon



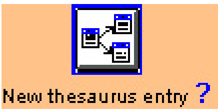
which results in :-

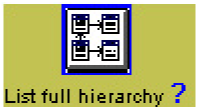


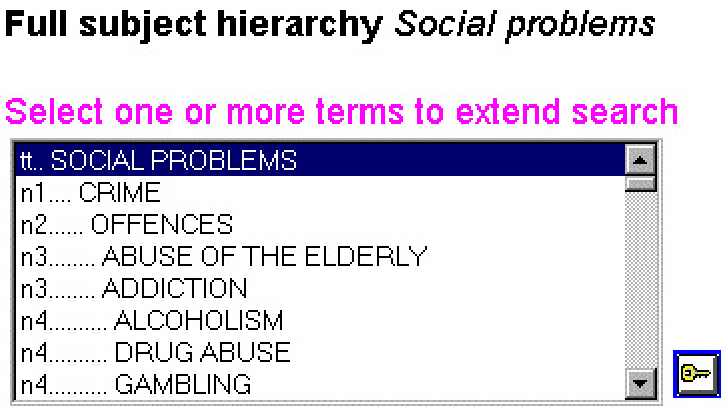
497 studies found for period ***1990*** to ***1997***
55 studies found for subject ***Alcoholism***, *Drinking Offences*
129 studies found for location ***Denmark***
5 unique studies found from integrated extended search:

Then from “alcoholism” you could select the top term “social problems”

and click on the followed by the to display the full hierarchy







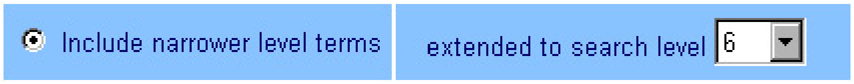
Full subject hierarchy *Social problems*

Select one or more terms to extend search

tt.. SOCIAL PROBLEMS
n1.... CRIME
n2...... OFFENCES
n3........ ABUSE OF THE ELDERLY
n3........ ADDICTION
n4.......... ALCOHOLISM
n4.......... DRUG ABUSE
n4.......... GAMBLING

and select up to ten terms from the listing of 269 terms. N.B. n4 indicates a narrower term 4 levels below the selected term. The maximum level for this hierarchy is n6.

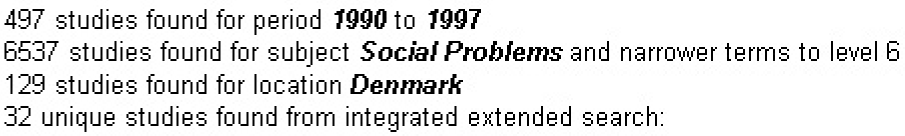
All 269 terms can be selected for a search by returning to the thesaurus listing and selecting the following options before clicking on the extend icon.



● Include narrower level terms

extended to search level 6

Which results in:-



497 studies found for period ***1990*** to ***1997***
6537 studies found for subject ***Social Problems*** and narrower terms to level 6
129 studies found for location ***Denmark***
32 unique studies found from integrated extended search:

FUTURE DEVELOPMENTS

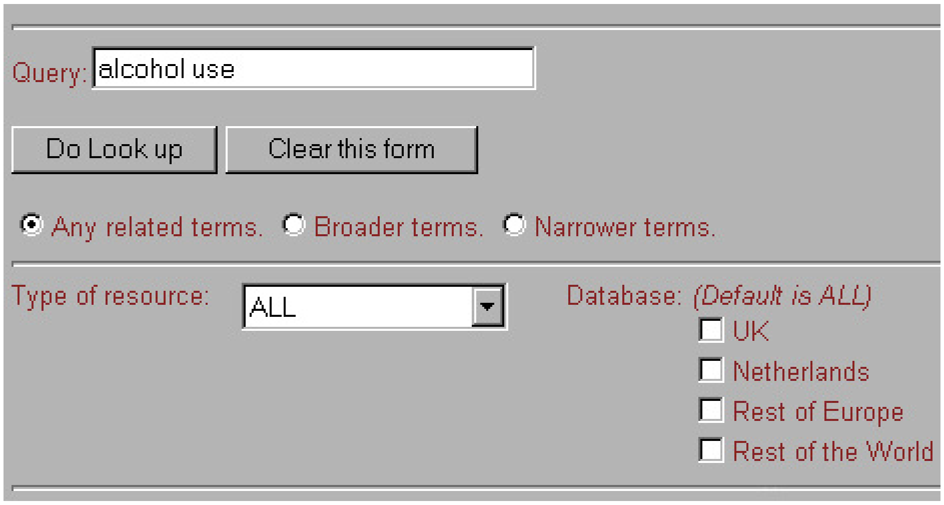
It must be remembered that HASSET has been constructed based on the 5,000 datasets held at The Data Archive as an indexing tool. So therefore the coverage only reflects the subject coverage of these datasets themselves, and because it is a controlled vocabulary it has not been specifically designed as a free text retrieval tool. However, its use within BIRON has seen an increase in the number of USE and UF relationships. So although the study descriptions and dataset documentation have not been trawled for candidate terms, which are then structured into a thesaurus, The Data Archive and several external organisations are experimenting with using HASSET as a retrieval tool for free text searching.

The CESSDA (Council for European Social Science Data Archives) IDC (Integrated Data Catalogue) is based on a Z39.50-WAIS protocol and uses freeWais-sf and SFgate as its search engine and gateway.

One of the options when creating an index for a WAIS database is to have present a synonym file, however since WAIS indexes every word, apart from stop words such as and, the etc., the synonyms have to be single words themselves. There is also no facility for the narrower / broader type relationships or control over when to apply the synonyms to a search. Hence we have selected only single word terms with a USE relationship to another single word term and single word top terms that have a RT relationship with other single word top terms. Part of the NESSTAR project will be to investigate how HASSET can be employed more fruitfully across the distributed databases of the European data archives and whether a multi-lingual version is a viable option.

Other organisations have also shown an interest in HASSET, namely SOSIG (Social Science Information Gateway), MIDAS (Manchester Information Datasets and Associated Services), QUALIDATA (Qualitative Data Archival Resource Centre), the Steinmetz Archive for the EU-funded ILSES project, IBSS (International Bibliography for the Social Sciences) and the Office for National Statistics in the UK.

The most advanced of these is SOSIG who have a test interface on the WWW which they hope to incorporate into their search facility by June 1997. They have matched terms in the HASSET thesaurus against keywords used in their own database records.



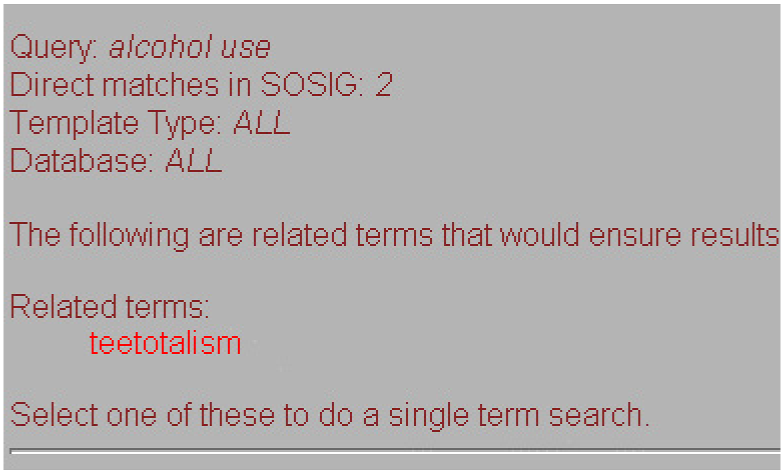
Query: alcohol use

● Any related terms. ○ Broader terms. ○ Narrower terms.

Type of resource: ALL

Database: *(Default is ALL)*
□ UK
□ Netherlands
□ Rest of Europe
□ Rest of the World

By keying in a search string and selecting the ‘Any related terms’ button and clicking on ‘Do Look up’ the present test interface will return the number of direct matches and also any term from the HASSET relationships that are guaranteed to result in a match in SOSIG.



Query: *alcohol use*
Direct matches in SOSIG: 2
Template Type: *ALL*
Database: *ALL*

The following are related terms that would ensure results

Related terms:
teetotalism

Select one of these to do a single term search.

The Data Archive hopes to undertake a project later this year where these participating organisations help convert HASSET into a thesaurus resource for the whole of the social science community. Control and maintenance of HASSET will still remain the responsibility of The Data Archive, but the other organisations will offer up candidate terms and suggestion position in the hierarchies through a new WWW interface to HASSET. As well The Data Archive will also review the contents and structure of HASSET through analysis of the search logs from BIRON and a trawl of the study descriptions and recently digitised dataset documentation. It is hoped that this will also make HASSET a valuable, universally available, free-text retrieval tool.

[*] Paper presented at IASSIST/IFDO ‘97, Odense, Denmark, May 6-9,1997. Ken Miller - Database programmer, The Data Archive, University of Essex, England.