Project ref. no.	*IST-1999-11748*
Project acronym	**LIMBER**
Project full title	**Language Independent Metadata Browsing of European Resources**

Security (distribution level)	*Public*
Contractual date of delivery	*M07 July 2000*
Actual date of delivery	*31 July 2000*
Deliverable number	*D-1*
Deliverable name	*Requirements Definition*
Type	*Report*
Status & version	*Final*
Number of pages	*42*
WP contributing to the deliverable	*WP3*
WP / Task responsible	*UKDA*
Other contributors	*NSD, Intrasoft, CLRC, User Group*
Author(s)	*Ken Miller, UKDA*
EC Project Officer	*Mr Kimmo Rossi*
Keywords	*System requirements, multi-lingual thesauri, search interfaces, metadata standards, usability, inter-operability*
Abstract (for dissemination)	*This requirements document outlines the needs of the potential user group community from the deliverables of the LIMBER project. As well as overall system requirements, including inter-operability with existing systems, specific details are included which cover the mono-lingual and multi-lingual thesauri, their use and maintenance, search interfaces, usability and metadata production, display and maintenance.*

Contents

Executive Summary

This deliverable defines the requirements on the LIMBER system to meet the needs of European users of data held in Social Science Data Archives to store and retrieve that data in order to plan and make policy decisions.

LIMBER will develop generic technology to internationalise thesaurus based access to metadata archives. LIMBER will be demonstrated by enhancing the existing NESSTAR system which will be improved in two EU funded IST projects, LIMBER and FASTER.

In LIMBER the three new requirements to be met are:

1) cross-domain barrier preventing integration of data from the social science archives with data from other sources will be tackled via the adoption of the World Wide Web Consortium (W3C) Resource Description Framework (RDF) or XML-schema for describing metadata - the choice is dependent on the DDI social science metadata standardisation initiative;
2) LIMBER will use a multi-lingual version of the thesaurus of the UK Data Archive, HASSET and multi-lingual interfaces to overcome the linguistic barrier;
3) assignment of terms from the multi-lingual thesaurus will be aided by the creation of an automatic indexing tool.

1. Introduction

The data archiving movement began in social science departments in the United States in the 1960s and rapidly spread to Europe: The Data Archive at the University of Essex was founded in 1967. The first data archives collected data of specific interest to quantitative researchers in the social sciences but more recently the needs of qualitative researchers have been recognised. Other initiatives have seen the development of a distributed service devoted to the archiving and dissemination of data across a broad range of disciplines and data types of interest to teachers and researchers in the arts and humanities.

In the context of data archives `data' means computer-readable data. They acquire, store and disseminate data for secondary research. This implies that the data collected for a primary purpose are then made available for research by other individuals or groups. This research may seek to replicate analyses already carried out by primary researchers in order to verify, extend, or elaborate upon the original results or to analyse the data from an entirely different perspective. Censuses and surveys carried out by governments for their own policy purposes are particularly rich sources of data for further exploration.

The original data need not necessarily have been collected specifically for research. Administrative databases, such as National Health Service Patient Re-registrations, show where patients are re-registered when they move from one Family Practitioner Area to another as part of a management information system. These data, although collected for a very different purpose, yield valuable and timely information for external researchers on migration patterns.

Data are created in a wide variety of formats. Data archives typically collect numeric data, which can then be analysed with the use of statistical software. Numeric data may result when textual information (such as answers to survey questions) has been coded or they may represent individual or aggregated quantities, for instance of sums of money earned or goods exported. Increasingly however as more flexible and powerful text retrieval and database management software become available, there is a demand for the archiving and dissemination of the computer-readable texts themselves. A variety of such materials from literary texts to anthropological notes to transcripts of historical source material as well as the complete answers to survey questions are now available for computer analysis.

Data archiving is a method of conserving very expensive resources and ensuring that their research potential is fully exploited. Unless preserved for further research, data which have often been collected at significant expense, with substantial expertise and involving respondents' time and effort may later exist only in a small number of reports which analyse only a fraction of the research potential of the data. Within a very short space of time the data files are likely to be lost or become obsolete as the technology of the host institution changes. Data archives store, catalogue, index and disseminate the data for further contemporary or future historical research.

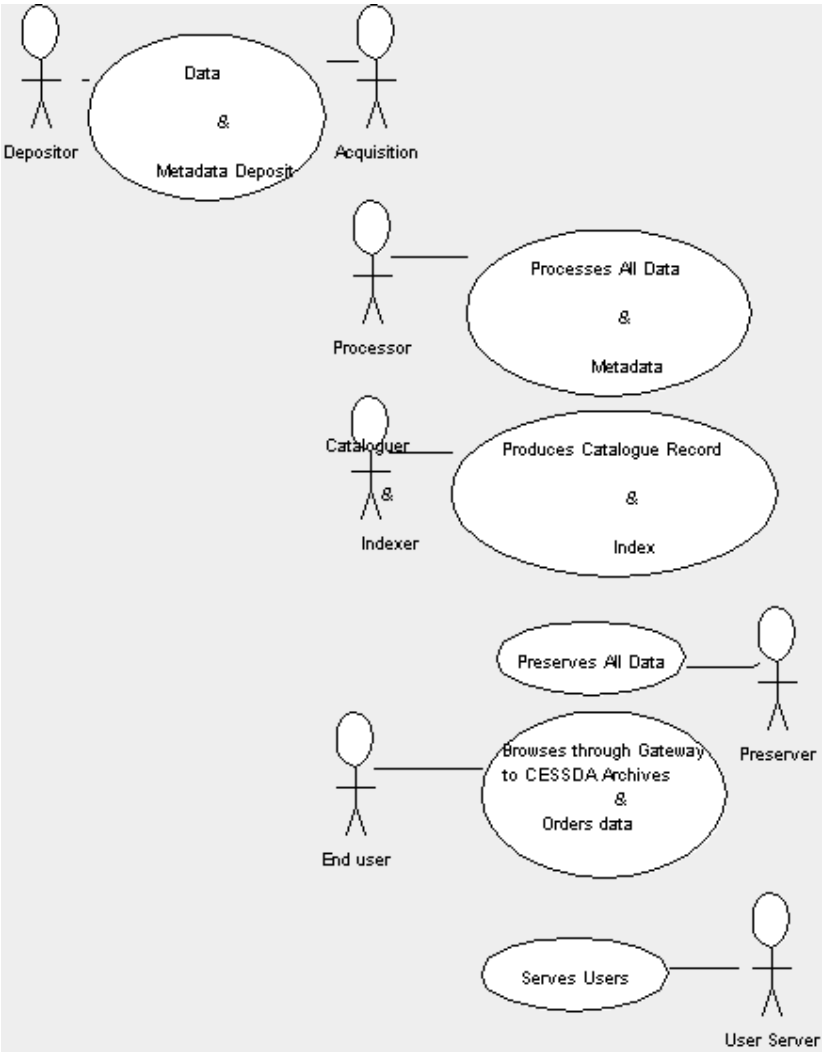
Archives ensure that when technology changes, the data in their holdings are technically transformed to remain readable in the new environment. To make this process easier they transform data on receipt to a standard in-house format from which they may be easily re-formatted in one sweep to the specifications required by changing technology. Suites of

programs are maintained to allow data to be easily transformed from the in-house standard to the various formats required by the computing environments in which individual users work.

Thus users of archives can be supplied with data in a format and on media which are appropriate to their needs. This allows researchers to conduct research in the computing environment with which they are familiar and which is most suitable for their research.

2. Data archive business process model

The Data Archive is an electronic archive that provides long-term preservation of and access to significant social science datasets. It is a dedicated online repository with the electronic objects contained within it maintained on an ongoing basis and converted to new formats as standards demand. The business process is summarised in the following diagram.



2.1. Overview

2.1.1. The Data Archive objectives

The Data Archive has, as its basic objectives to:

- identify, select, catalogue, preserve and disseminate significant social science datasets
- make the information available to users in line with depositors conditions of use
- take into account the depositor's commercial interests with regard to provision of access to remote users
- update the information in the archive on an ongoing and systematic basis, while maintaining date stamped previous 'editions'
- convert datasets in the archive to new formats as standards change.

2.1.2. The service model and business process modelling

This business process model is based on the current operational service model.

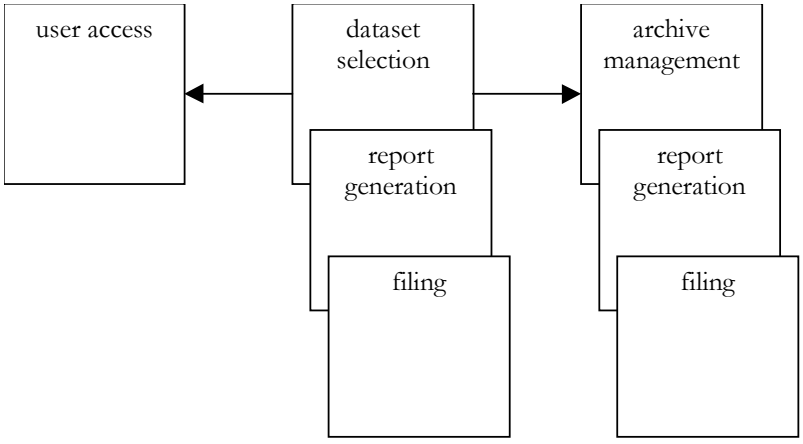
2.1.3. Copyright

The copyright of depositor's is recognised and displayed for all of the holdings in the archive.

2.1.4. Components of the Business Process Model

The Business Process Model is divided into *5* components, identified as:

- the user access component - disseminate
- the dataset selection component - identify and select
- the archive management component - catalogue, preserve and copyright
- the report generation component - monitoring and performance indicators
- the filing infrastructure - management of correspondence files.



These components comprise the archive's business processes.

2.2. The user access component

This component comprises functions relating to users gaining access to the archive.

2.2.1. Users

The users have been identified as:

- government departments, policy makers, private enterprise, the curious public, serious researchers, depositors, other archives and internal users.

2.2.2. User entry points

Users enter the Archive through the web site, catalogue search interfaces and through direct contact with our User Services section by mail, e-mail, fax and telephone.

- Browsing
- Free text
- Via thesaurus controlled vocabulary
- Verbal

2.2.3. Internal User Needs

- Technical Services need access in order to obtain information in relation to the management of the electronic archive
- Depositor Services need access in order to maintain information for the archive's collection
- Preservation Services need access to make decisions about preserving the physical manifestation of each electronic dataset
- User Services need access in order to identify and disseminate datasets

2.2.4. Guidelines for user access

Guidelines cover the following areas:

- datasets which are not commercially viable are accessible free of charge (except for media costs) to both internal and external users
- any depositor-imposed restriction on access to a dataset is negotiated with the depositor to achieve a standardised approach to access restrictions in the archive.
- any restriction imposed on access to a dataset is described in its title page

Negotiations are conducted with depositors to achieve standard conditions of access. If individual depositors cannot agree to the Archive's conditions, their material may be excluded from the archive. This will depend in part on the nature of the depositor's conditions and the importance of the dataset.

2.2.5. Dissemination

The Archive disseminates datasets via differing media types; principally CD-ROM and various forms of magnetic media. Datasets are also made available by ftp and metadata is accessible through web browsers and PDF files through Adobe Acrobat reader.

2.3. The selection component

2.3.1. The identification process

The Acquisitions section regularly scans available information sources to identify datasets in scope for preservation by the Archive. Suggestions also come from users and other members of staff.

When a dataset is identified, the first task is to check the dataset's details to determine suitability for archiving. The process of checking the details will result in highlighting the dataset as having potential for preservation, or rejecting the dataset for preservation. The results of checking the details are recorded to ensure that the status of any dataset is known at any time.

2.3.2. The selection process

Acquisitions staff assess each dataset, there are three possible outcomes:

- the dataset is accepted for archiving
- the dataset is rejected
- the dataset does not contain sufficient information to make a decision. It is placed on hold while the depositor is contacted.

The appropriate decision is recorded. Where a dataset is rejected, the reasons for its rejection are noted.

Contact with data creators

If the depositor does not respond, then the dataset is filed for future follow-up action. Any communications between the depositor and The Data Archive are filed in the Archive's registry file system.

2.3.3. The technical evaluation process

After a dataset has been identified and selected for preservation within the Archive, a technical evaluation is carried out.

- The documentation is checked against the data
- Results in final report checked
- The frequency of major variables checked
- Variable and value labels checked
- File size recorded
- The most appropriate file structure for the object and its versions is determined (see section 5 for more detail)

2.4. The archive management component

The major activities of archive management fall under the following headings:

- adding new datasets
- enabling access
- updating
- deleting

2.4.1. Addition of new datasets to the archive

The initial tasks required of the archive manager are to:

- create a catalogue record
- file all related correspondence (see section 7 for more detail)
- collect the depositor's metadata (see section 5 for more detail)
- create other archive metadata
- create the title page
- record contact details

If the dataset is selected for preservation, the following actions occur:

Allocation of unique identifiers

Each dataset is given a unique study number to track that dataset through its lifecycle within the Archive and for cataloguing and retrieval purposes.

Recording descriptive metadata

Descriptive metadata are noted for the dataset. These include:

- title, depositor/producer, abstract, subject category, access conditions, copyright provisions, methodology and assigning keywords

Cataloguing is based on the AACR2 guidelines.
A title page is created by Information Section staff for each dataset

archive title page

For each dataset, the following information is to be displayed where appropriate:

- History and provenance, copyright and introductory information

2.4.2. Enabling access

Tasks allowing the datasets in the archive to be made accessible to the user include:

- adding a new dataset to the search engine's indexes

- converting datasets from non-standard to standard data formats if required (see section 5 for more details)
- setting the dataset's access profile.

2.4.3. Updating

The updating tasks form an ongoing workload for the Information Section staff. These tasks may include:

- adding new "editions" or versions to dataset
- adding new releases to a serial dataset
- updating the metadata to reflect the addition of new content or other information about the object
- updating the title page details.
- contact details

The updating tasks are necessary to assure the quality and completeness of the content and framework of each dataset.

The gathering schedule

The gathering schedule, or time interval between each successive capture, must be managed in its own right.
The gathering schedule shows the date of capture for each electronic document and should give details of frequency, for example daily, weekly, monthly and the date of last gathering. The gathering schedule does not necessarily match the date of issue of any electronic dataset or its objects.

2.4.4. Deleting

Defined as removing whole datasets from the archive. Associated tasks include:

- removing all content from the archive
- updating the metadata.

2.5. Issues of archive management

2.5.1. Standards for metadata

Information describing the electronic datasets, (which will allow easy access to them and management of them in the archive) needs to be captured as smoothly as possible. One means of attaining this is by using a standard set of metadata. This information is managed under the Data Documentation Initiative's (DDI) study description and codebook standard with a mapping to Dublin Core metadata standards. Using these standards gives the potential for automated capture of the information.

2.5.2. Administrative metadata

Administrative and preservation metadata may need to be captured at any time during the maintenance of a digital object. This metadata is in addition to any required under DDI or Dublin Core elements.

2.5.3. Standard storage formats

Digital object formats support full text, images, text as image, audio and video. There is not one particular standard format that presides over the others at the moment. However, the two major contenders for text based information are the standard storage formats of PDF and SGML - like formats. Depositors are likely to reflect the breadth of choice available by selecting formats that meet their strategic needs and budgets. In addition, they may use several formats within the one dataset. Where a depositor creates a dataset in a format, which is deemed to be non-standard, options for the Archive are to:

- negotiate with the depositor to provide a copy of the dataset in a standard format
- convert the dataset to one of the archive's standard formats.

2.5.4. File structure for the digital objects within the archive

The structure of digital objects varies from depositor to depositor. Each depositor creates a unique set of file naming conventions to bring together the disparate parts of a dataset. The archive must achieve two goals with its file structure:

- provide correct access to datasets at all times
- deal with inconsistencies between depositors' file structures

2.6. Report generation component

Two sets of users will require reports from the management facility:

- managers, who will require daily activities to be monitored and reported
- other interested parties such as funders and depositors.

Reports are generated to provide information on many areas including:

- the status of the archive, for example the number of new archived datasets in a given timeframe, the number of datasets which have had information changed, the overall number of datasets, the backlog of datasets which have been captured but linked to access points
- preservation which must contain the metadata to allow decision-making to occur regarding the preservation action to be taken. Analysis by formats, age, software used etc by dataset will be required
- technical reports such as; time taken by the capture process; total space usage by archive; traditional database software reports; transaction log activity, including counts of the number of times each dataset has been accessed (may be analysed further by location of browser; comparing the pre- and post-migration version of any object where there has been a change in standard.

2.7. The filing infrastructure component

There is on going communication between the Archive and the depositor that needs to be captured as a true record of digital object creation. This is essential for future action and evidentiary purposes.

GLOSSARY of terms as used in the Business Process Model

Term	Definition
Archiving	the process of acquiring datasets, storing them and managing access to them within the archive. *See also* Preservation.
Dataset	A dataset is a manifestation of a digital object. The dataset can be a single digital object or it may consist of a hierarchical structure of digital objects. Lower level objects (such as graphics and sounds) may be stored with the object (which then remains a single electronic file, for example, a MS Word document) or as separate objects (for example, a Web page, which stores graphics separately).
Preservation	the process of initiating strategies and undertaking activities to ensure that archived datasets will remain viable and accessible in the long term. *See also* Archiving.

3. Scenarios

3.1. General Data Archive Scenarios

In the overall process of archiving and retrieving data there are four major scenarios:

1) Archive a dataset
2) Publish a dataset
3) Retrieve a dataset from archive
4) Supply a dataset

The actors involved in these and the conditions that must be met for goal completion are:

Actor	Conditions	Goal
Data Preservation Section	CHECK DATA CHECK DOCUMENTATION CONVERT TO STANDARD FORMAT DATASET MOVED TO PRESERVATION AREA	Dataset archived
Information Section	CREATE CATALOGUE RECORD ADD INDEX TERMS FROM THESAURUS MAKE AVAILABLE IN SEARCH DATABASE LINK ADDITIONAL PDF DOCUMENTATION	Dataset published
User	SEARCH RESOURCES	Obtain dataset

	DETERMINE RELEVANCE ORDER DATASET RECEIVE DATASET	
User Services Section	INTERPRET ORDER CONVERT DATA TO REQUIRED FORMAT ATTACH RELEVANT DOCUMENTATION DISSEMINATE DATASET	Supply dataset

3.2. Overview

European data archives currently archive data, indexed by metadata, accessible through the NESSTAR system that includes a client and server programs.

The Data Documentation Initiative (DDI) XML Codebook Document Type Definition (DTD) is an internationally defined standard for social science metadata. It provides a comprehensive set of elements, including those to describe the scope and dimensions of the data, temporal and geographic coverage, methodological information, variables and questions text. CESSDA (Council for European Social Science Data Archives) members have adopted the standard allowing inter-operability within the NESSTAR system (Networked European Social Science tools and Resources). Each European social science data archive created a Z39.50 database of their holdings prior to the DDI standard, and subsequently indexed the metadata from various elements of the DDI DTD to create 36 common indexes. Although the tags of each element are common (abbreviated English), the information held between them are in the various languages of the members country and the only resources available are those of the CESSDA archives.

Currently two EU funded projects are developing the NESSTAR system further, this project LIMBER and the FASTER project. Each will address different requirements to improve the access to the archived data.

In LIMBER the three new requirements to be met are:

4) cross-domain barrier will be tackled via the adoption of the World Wide Web Consortium (W3C) Resource Description Framework (RDF) or XML-schema standards for describing metadata;
5) LIMBER will use a multi-lingual version of the thesaurus of the UK Data Archive, HASSET and multi-lingual interfaces to overcome the linguistic barrier;
6) assignment of terms from the multi-lingual thesaurus will be aided by the creation of an automatic indexing tool.

A further major requirement on the LIMBER project is that the existing NESSTAR system, and the new FASTER system should each be able to be used without commitment to any additions resulting from LIMBER. This is to ensure that no IPR issues resulting from LIMBER prevent the continued use of NESSTAR and FASTER by the user population. Consequently, the three requirements above should be met by writing add-ons to the existing NESSTAR system, and not by significantly modifying it. The FASTER and NESSTAR consortium members can be asked to minimally

modify NESSTAR code so that they, and in turn the NESSTAR consortium retains IPR.

3.2.1. Present Indexing Procedure

Primary Actor		Goal
Indexer	EXAMINE DEPOSITORS FORM EXAMINE CATALOGUE RECORD SELECT TERMS FROM THESAURUS WHICH REFLECT OVERALL CONCEPTS EXAMINE QUESTIONAIRRE, VARIABLE LIST AND FINAL REPORT SELECT TERMS FROM THESAURUS WHICH REFLECT SPECIFIC CONCEPTS	Index dataset

Conditions: -

1. Concept not in thesaurus
   a. Consult existing thesauri
   b. Insert new concept with links to existing concepts and any broader concepts and lead-in terms required.
2. Concept used as a lead-in term to a broader concept
   a. Review previous usage of broader concept
   b. Convert concept to a narrower relationship
   c. Re-index those previously using broader concept
3. Concept in thesaurus does not reflect modern terminology
   a. Replace old concept term with new
   b. Convert old term to lead-in term
4. Concept chosen has no or little previous use
   a. Review previous usage
   b. Review similar datasets
   c. Revise decision if necessary

3.3. Overall System

The LIMBER system should have platform independent client software, be modular in design and use open standards to allow maximum interoperability with existing systems. All technologies used should be Web-aware or able to be easily integrated in a Web-environment. The interfaces between modules should be clean and well documented to allow for easy swapping of components.

Although designed specifically around a project constructed multilingual thesaurus it should be able to interact with other thesauri freely available in compatible formats and registered with a thesaurus-mapping registry. Although initially restricted to the four European languages, English, French, Spanish and German, the ambition of the project is to expand to included further European and other languages. Hence the system will have to be UNICODE compliant in order to deal with the various character sets that the extended scope would entail.

The server software needs to maintain and interface with metadata and thesauri marked up in RDF/XML, but avoiding the use of costly third-party components, which will prevent the exploitation of the technology.

NESSTAR, with the DDI codebook metadata standard, is an obvious resource for LIMBER to interact with. However, other resources within the social sciences, such

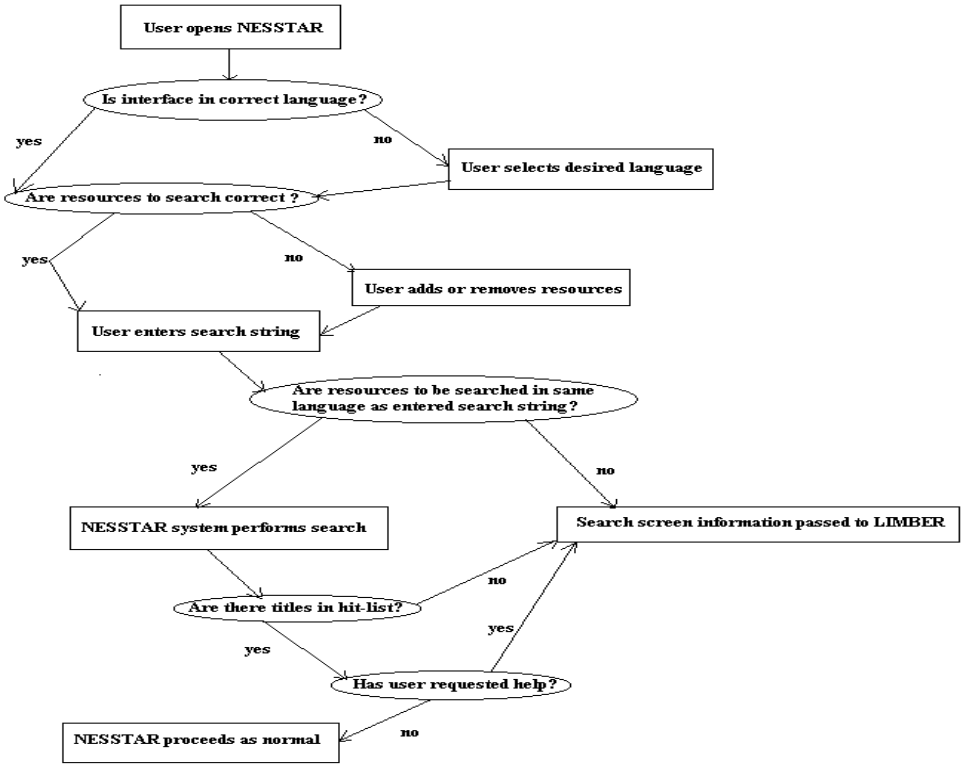
as SOSIG (Social Science Information Gateway), Qualidata (ESRC Qualitative Data Archival Resource Centre), CASS (Centre for Applied Social Surveys) Question Bank and REGARD (national database of ESRC funded research), should also be able to inter-operate with the LIMBER system.

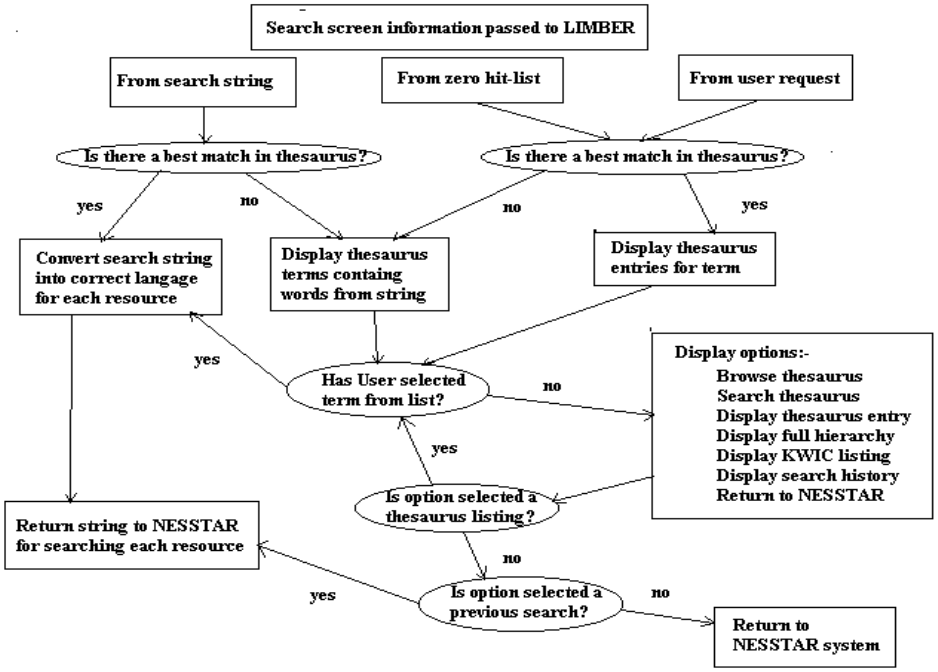
Although, all but one of the above uses the HASSET thesaurus to index their resources, they all apply different metadata standards. Similarly resources from different domains will employ different metadata standards and also different thesauri.

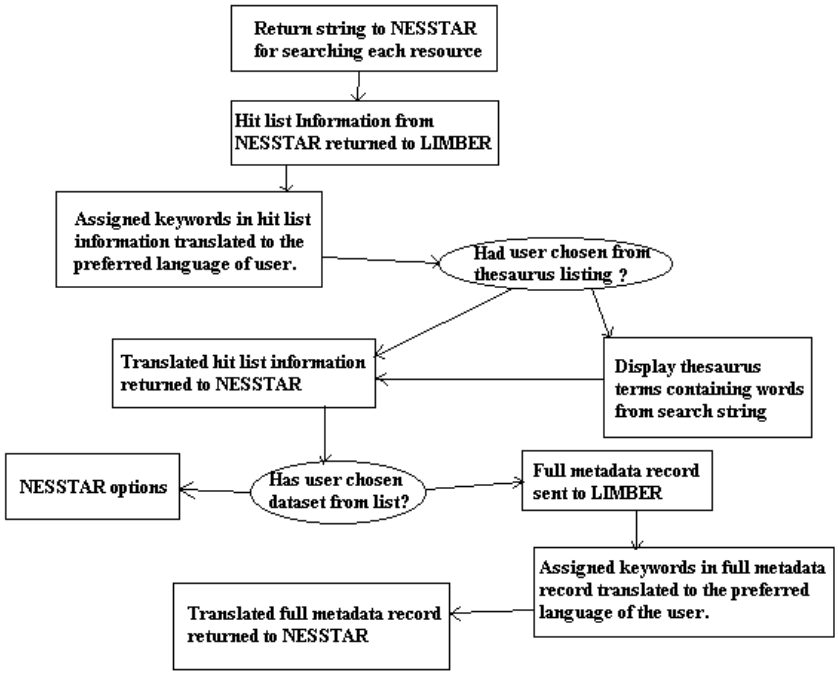
It is here that the work of NKOS (Networked Knowledge Organisation Systems) and SCHEMAS (Forum for Metadata Schema Implementers) should be closely monitored. NKOS, through their thesaurus registry and thesauri mappings, are trying to build Internet discovery architecture among services based on subject indexing languages or domain ontologies. SCHEMAS are planning the same for metadata schemas.

3.3.1. LIMBER / NESSTAR interaction

Below are shown the interactions of users with the LIMBER/NESSTAR system.




Search screen information passed to LIMBER
From search string
From zero hit-list
From user request
Is there a best match in thesaurus?
Is there a best match in thesaurus?
yes
no
no
yes
Convert search string into correct langage for each resource
Display thesaurus terms containg words from string
Display thesaurus entries for term
yes
Has User selected term from list?
no
Display options:-
Browse thesaurus
Search thesaurus
Display thesaurus entry
Display full hierarchy
Display KWIC listing
Display search history
Return to NESSTAR
yes
Is option selected a thesaurus listing?
Return string to NESSTAR for searching each resource
no
yes
Is option selected a previous search?
no
Return to NESSTAR system


Return string to NESSTAR for searching each resource
Hit list Information from NESSTAR returned to LIMBER
Assigned keywords in hit list information translated to the preferred language of user.
Had user chosen from thesaurus listing ?
Translated hit list information returned to NESSTAR
Display thesaurus terms containing words from search string
NESSTAR options
Has user chosen dataset from list?
Full metadata record sent to LIMBER
Assigned keywords in full metadata record translated to the preferred language of the user.
Translated full metadata record returned to NESSTAR

4. Detailed Requirements

4.1. Monolingual Thesaurus

The reduction of the existing HASSET thesaurus should be based on removal of cultural and institutional specificity and the existence of other domain thesauri. A feedback mechanism for the appraisal of reductions and additions from each site should be set up. **(3.1)**

The monolingual thesaurus should also include hierarchies to cover the elements of the DDI codeBook standard that would aid determining compatibility between datasets, such as methodology, kind of data, universe, spatial unit, access conditions and file structure. Wherever possible these should build on existing listings. **(3.2)**

4.2. Multilingual Thesaurus

Specialist teams, with social science backgrounds, at each user group site should oversee the translation of the monolingual thesaurus. A mechanism, such as an email discussion group, should be set up for cross appraisal of translations and a further mechanism to deal with the possible translations of new additions of synonyms from each site. **(3.3)**

The multilingual thesaurus should be designed to allow different hierarchical structures in each language and non-equivalence between the terms of each language. The multilingual thesaurus should employ widespread use of scope notes, to aid searching and definition of hierarchies, terms and non-equivalence. Version control is essential to ensure compatibility and migration paths as the thesaurus expands and adapts. **(3.4)**

4.3. Metadata

Limber will provide a new metadata system to support resource discovery.

The Metadata system will provide at least the same functionality as the existing searching system within Nesstar. Thus it will:

1 Support metadata manual indexing as currently supported in Nesstar.
2 Support metadata searching and retrieval, similar to the existing Cheshire system within Nesstar.
3 Conform to the API within the existing Nesstar system in order to communicate with other tools in the Nesstar system.
4 Existing metadata entered in the system can be handled by the Limber metadata system. The current metadata uses an XML format; conversion will be required to and from the proposed RDF format.

The metadata system will support multi-linguality. Thus it will:
1 Support multiple language keywords in metadata fields (one resource, one description multiple language entries)
2 Supply language fields in the metadata
3 Use the thesaurus to support multi-language keyword indexing.
4 Use the thesaurus to look up multi-language alternatives to search terms and to terms in the metadata.

The metadata system will be designed to support multiple metadata formats, allowing cross-domain searches. Dublin Core is an initial candidate, allowing searches into bibliographical databases.

The metadata format will be compatible with the metadata format developed in the Faster project. Faster will concentrate on providing metadata support for “data cubes” and statistical data.

The metadata developments within Limber should support the further progress and investigations of the DDI initiative. For example:
1 Investigate the use of RDF to capture the DDI;
2 Support the use of XML Schema to support the DDI;
3 Extend the DDI with multilingual support;
4 Extend the DDI as appropriate in conjunction with the Faster project.

4.4. Usability

Quantitative and qualitative results should be collected from the user group to identify and correct problems with the functionality and usability. After the usability testing, an evaluator will judge which of ten usability heuristics (developed by Molich and Nielsen, 1990 and refined by Nielsen, 1994) each problem recorded in the user testing breaches.

1. The system should always keep users informed about what is going on, through appropriate feedback within reasonable time.

2. The system should speak the users' language, with words, phrases and concepts familiar to the user, rather than system-oriented terms. Follow real-world conventions, making information appear in a natural and logical order.

3. Users often choose system functions by mistake and will need a clearly marked "emergency exit" to leave the unwanted state without having to go through an extended dialogue. Support "undo" and "redo".

4. Users should not have to wonder whether different words, situations, or actions mean the same thing. Follow platform conventions.

5. Even better than good error messages is a careful design, which prevents a problem from occurring in the first place.

6. Make objects, actions, and options visible. The user should not have to remember information from one part of the dialogue to another. Instructions for use of the system should be visible or easily retrievable whenever appropriate.

7. Accelerators -- unseen by the novice user -- may often speed up the interaction for the expert user such that the system can cater to both inexperienced and experienced users. Allow users to tailor frequent actions.

8. Dialogues should not contain information, which is irrelevant or rarely needed. Every extra unit of information in a dialogue competes with the relevant units of information and diminishes their relative visibility.

9. Error messages should be expressed in plain language (no codes), precisely indicate the problem, and constructively suggest a solution.

10. Even though it is better if the system can be used without documentation, it may be necessary to provide help and documentation. Any such information should be easy to search, focused on the user's task, list concrete steps to be carried out, and not be too large.

Copied from http://www.useit.com/papers/heuristic/heuristic_list.html

4.5. Search Interface

The LIMBER interface should be available in all languages and be simple, well designed and well documented. Special care should be taken with the explanation of how words are combined by default in a search and how those defaults can be overridden. An option should be available to switch between a free-text and a keyword search. The default for all searches should be across all sites and in all languages and wherever possible return a hit list of datasets, automatically substituting synonyms with the preferred thesaurus term. The free text search option should employ translation and synonyms, whereas the keyword search should employ the assigned concept classification code. This would increase search speed, and allow pick up of narrower local variations from the central authoritative version of the multilingual thesaurus. The use of stemming, truncation or fuzzy searching should be employed on words from a failed search string to list the terms from the thesaurus which contain those words or derivations. **(3.5)**

The search should be fast and retrieve relevant resources ranked in the importance defined by the user. The ranking should be based on the number of occurrences and position of search terms, geography and time, with an explanation available to show why any given dataset has been included. **(3.6)**

All searches should be saved in a search history with the option to combine, using Boolean operators, with other saved searches or new searches. The hit list should be accompanied by a suggestion list from the controlled vocabulary of the thesaurus. The complexity of the thesaurus should only be displayed when requested by the user or in a browsing interface. Otherwise a simple alphabetic listing of the relationships

to the selected thesaurus term should be displayed. The browsing interface should provide the user with the ability to drill down into any hierarchy of the multilingual thesaurus starting from its top term. **(3.7)**

4.6. Relevance Feedback

If available, the LIMBER system should display the titles of retrieved datasets in the user's language. The languages in which metadata is recorded should be available from the hit list with the ability to select the language in which to display the full metadata record. Automatic translation of some textual elements could be tried, otherwise most resources will have a title and/or summary in English as an alternative to their natural language. The use of relevance feedback from the translation of assigned keywords should only be on the specific elements of the DDI codeBook standard, identified in the thesaurus work package, that would aid determining compatibility between datasets. It is essential that the names and definitions of all elements from the DDI codeBook standard be translated. The LIMBER project should co-ordinate with data creators and originating archives to offer further external assistance in cross-national analysis, such as links to translated versions of documents and translation services. **(3.8)**

4.7. Thesaurus/Metadata Servers and Tools

Each site will require a structured definition of their resources, the languages covered and any imbalances in metadata and assigned keywords, so that any multilingual, multi-site search can be rationalised and speeded up. **(3.9)**

The multilingual thesaurus will be marked up in either XML or RDF so maintenance will have to be via a XML or RDF compliant database. The server will have to return the classification code, language equivalents or relationships from an exact match of a concept from the multilingual thesaurus. Whole hierarchies need to be extracted or constructed for the browsing interface. Also concepts containing words or stems or truncations from either another concept or search string have to be extracted. **(3.10)**

For the automatic assigning of keywords to totally duplicate manual procedures, the tool would have to work from depositor's forms and the original data producer's codebook or questionnaire. Examples of the UKDA versions of these documents, and a marked-up UKDA DDI metadata record are supplied in the Appendix. **(5.)**

However, each resource will have different sources for this type of material. Perhaps a more realistic tool could work from the actual metadata marked up to the DDI codeBook standard. Here assigned keywords could be converted to the LIMBER multilingual thesaurus equivalent and classification codes assigned. In the specific areas of the DDI codeBook standard identified in the thesaurus work package free text could be scanned to assign keywords and codes from the very limited hierarchy of terms. Although more difficult, similar techniques could be employed on more general areas such as title, abstract, variable label and question text. The tool should

also convert the XML metadata from DDI version 1.0 to the new version required for LIMBER. **(3.11)**

Any thesaurus management system used should perform as well as the commercially available competition (see Appendix section 8.5), on each of the usual criteria for their performance (see Appendix section 8.4).

4.7.1. Thesaurus Displays

N.B UF - Use For
BT - Broader Term
TT - Top Term
RT - Related Term
(Also used in standard thesaurus display:- NT - Narrower Term and USE - Use)

Standard

INTERPERSONAL ATTRACTION P80.30/90
UF COURTSHIP
FRIENDSHIP
LOVE
BT INTERPERSONAL RELATIONS >>
TT BEHAVIOUR >>
RT FRIENDS

Hierarchical

BEHAVIOUR
HUMAN BEHAVIOUR
SOCIAL BEHAVIOUR
SOCIAL INTERACTION
INTERPERSONAL RELATIONS
INTERPERSONAL ATTRACTION **
COURTSHIP
FRIENDSHIP
LOVE

KWIC (Keyword in context)

INTERPERSONAL ATTRACTION **
INTERPERSONAL COMMUNICATION
INTERPERSONAL CONFLICT
INTERPERSONAL INFLUENCE
INTERPERSONAL INTERACTION
INTERPERSONAL RELATIONS

4.8. Requirements for integration with NESSTAR

The principle of integration with NESSTAR is that NESSTAR must be able to run without any changes after LIMBER has been developed. The intention is that NESSTAR should operate in conjunction with modules developed to in LIMBER to meet the requirements stated in this document.

NESSTAR can be viewed as a three layer architecture:

1) Data and metadata servers
2) NESSTAR server
3) NESSTAR client

R1) There should be no changes required to the NESSTAR server layer architecture or implementation at the second level in order to meet the LIMBER requirements.

Conformance Test for R1) – Run the NESSTAR client and server with and without the LIMBER components, without any changes required to the server.

R2) The NESSTAR client contains data files for the user interaction strings used in menus, error messages etc.. This data file should be populated with strings for each of the languages addressed by LIMBER to provide a multi-lingual interface.

R3) The NESSTAR client user preferences should support the use statement and use of strings to state the language preferred by the user from among the set supported by LIMBER.

Conformance Test for R2 and R3) – run through the interaction dialogues shown in this requirements document in each of the languages addressed in LIMBER to show that each language is fully supported.

R4) The functionality required to browse the thesaurus should be provided in such a way (e.g. a separate window) that it can be launched from the NESSTAR client, a query can be constructed and saved in a file, then the query can be launched from NESSTAR without altering the NESSTAR client.

R5) The changes to the metadata representation required in LIMBER can allow the re-coding of the metadata currently stored in NESSTAR in the CHESHIRE data tool, both with a new data representation and with a new storage tool. However, the calls to and from this metadata store made by the NESSTAR server should still be supported in details of their storage and retrieval functionality, and in the details of the access protocols used.

Conformance test for R5) – the NESSTAR server can perform its functions on any new metadata representations and stores.

R6) Tools developed to meet the requirement for a semi-automatic indexing tool for metadata will produce data that can be stored in the new metadata repository. These tools should not interact directly with the NESSTAR architecture.

Conformance test for R6) – stand alone demonstration of tools developed to meet the requirement for a semi-automatic indexing tool for metadata.

R7) Tools developed to store and serve the new thesaurus representation should operate independently of the existing NESSTAR server layer and should not require any changes to it. Interaction required with the thesaurus server from the client should use either:

1) existing calls, protocols and formats from the NESSTAR client and server layers;
2) new calls from a new thesaurus browsing tool to the thesaurus directly.

Conformance test for R7) – demonstration of thesaurus use with the existing NESSTAR system.

5. Use Cases

5.1. Reduction of Monolingual Thesaurus

Primary Actor		Goal
Thesaurus Manager	REDUCE HASSET THESAURUS Conditions: - 1. Thesaurus too British a. Remove cultural specificity 2. Thesaurus reflects UKDA holdings a. Remove institutional specificity 3. Thesaurus not domain specific a. Remove hierarchies covered by existing thesauri 4. Reduction not suitable for other site a. Re-appraise reduction b. Adapt reduction and re-circulate 5. Additional terms required from other site a. Re-appraise structure **b.** Add terms and re-circulate	Broad based monolingual social science thesaurus

5.2. Addition of Specific Hierarchies

Primary Actor		Goal
Thesaurus Manager	ADD HIERARCHIES TO THESAURUS Conditions: - 1. Thesaurus not suitable for dataset comparison a. Add methodology hierarchy consulting existing lists b. Add kind of data hierarchy consulting existing lists c. Add universe hierarchy consulting existing lists d. Add spatial unit hierarchy consulting existing lists e. Add access conditions hierarchy consulting existing lists **f.** Add file structure hierarchy consulting existing lists	Extended monolingual social science thesaurus

5.3. Translation of Thesaurus

Primary Actor		Goal
Thesaurus Manager	TRANSLATE THESAURUS Conditions: - 1. Translated thesaurus not correct a. Set up specialist teams at each site b. Cross-site appraisal of translations c. Addition of language specific synonyms d. Translation of added synonyms	Multilingual equivalent social science thesaurus

5.4. Adaptation of Thesaurus

Primary Actor		Goal
Thesaurus Manager	ADAPTATION OF THESAURUS Conditions: - 1. Structure of translated thesaurus not correct a. Addition of language specific terms b. Allow one to many mappings of terms in different languages c. Add scope notes to describe hierarchies d. Add scope notes to explain ambiguity e. Add scope notes to explain non-equivalence f. Employ version control to ensure compatibility	Extended Multilingual equivalent social science thesaurus

5.5. Enter Search String

Primary Actor		Goal
User	ENTER SEARCH STRING Conditions: - 1. Search interface in wrong language a. Switch to native language interface 2. Search interface too complicated a. Switch to simpler interface 3. Combination of words in search string wrong a. Switch between free text and keyword option b. Click button for syntax help on combining 4. Concept not found a. Automatic checking of synonyms and switching to preferred concept b. Thesaurus help - alphabetic listing of all concepts containing words from string; employing stemming, truncation and fuzzy matching 5. Search only in natural language a. Free-text search with all synonyms and translations b. Keyword search on classification code	Dataset Hit List

5.6. Review Hit List

Primary Actor		Goal
User	REVIEW HIT LIST Conditions: -	Select Dataset(s)

1. Hit list titles and interface in wrong language
   a. Switch to native language interface
2. Hit list interface too complicated
   a. Customise hit list components
3. Presence of datasets in hit list not understood
   a. Click on hit list item to reveal explanation of inclusion
4. Order of datasets in hit list not understood
   a. Click button for explanation of ranking
   b. Customise ranking

5.7. Refine Search

Primary Actor		Goal
User	REFINE SEARCH Conditions: - 1. Retrieval not relevant a. Thesaurus help - alphabetic listing of all NT, BT & RT concepts and concepts containing words from original concept b. Switch to browsing interface - select concept from listing of top terms and drill down hierarchies until exact match found 2. Retrieval too large a. Customise ranking b. Apply temporal and/or geographic limits c. Save search and Boolean "AND"/"NOT" combine with another concept d. Save search and combine with previous other search(es) e. Return to previous search f. Thesaurus help - alphabetic listing of all NT, RT & BT concepts and concepts containing words from original concept 3. Retrieval too small a. Save search and Boolean "OR" combine with another concept b. Save search and combine with previous other search(es) c. Return to previous search d. Thesaurus help - alphabetic listing of all BT, RT & NT concepts and concepts containing words from original concept e. Remove temporal and/or geographic limits f. Add more resource sites to search	Different Dataset Hit List

5.8. Review Documents

Primary Actor		Goal
User	REVIEW FULL CATALOGUE RECORD REVIEW ACCOMPANYING DOCUMENTS Conditions: - 1. Catalogue headings and interface in wrong language a. Switch to native language interface	Confirm relevance

2. Catalogue metadata in wrong language
   a. Select alternative metadata elements in native language
   b. Display assigned keywords in native language
   c. Check for reference to external versions in native language
   d. Return to hit list and customise to show language of metadata
   e. Click help for translation services available
3. Catalogue interface too complicated
   a. Customise catalogue components

5.9. Review Sites

Primary Actor		Goal
System	REVIEW SITE DIRECTORY Conditions: - 1. Resources not relevant a. Ignore site in search 2. Languages not relevant a. Ignore site in search 3. Not all languages relevant a. Customise search for language 4. Imbalance recorded a. Customise search for metadata and/or indexing imbalance	Rationalise search

5.10. Update Thesaurus

Primary Actor		Goal
Thesaurus Manager	UPDATE THESAURUS Conditions: - 1. Resource corrupted a. Check indexes available for - classification code, terms, words, language equivalents, relationships and hierarchies b. Check XML/RDF compliance of records c. Check stemming algorithm correct d. Check reciprocal NT/BT relationships	Consistent, searchable resource

5.11. Index Metadata

Primary Actor		Goal
Indexing Tool	INDEX METADATA Conditions: - 1. Valid keyword assigned but without code a. Assign classification code 2. Keyword assigned not in thesaurus a. Convert to controlled vocabulary and assign classification code b. Update thesaurus - possible synonym or new term	Controlled vocabulary assigned LIMBER metadata

c. Software learns from conversion

3. Text in specific DDI elements not controlled
   a. Convert to controlled vocabulary and assign classification code
   b. Update thesaurus - possible synonym or new term
   c. Software learns from conversion
4. Text in general elements of DDI not assigned keywords
   a. Extract concepts from free text entries
   b. Map concepts to the controlled vocabulary
   c. Assign keyword and classification code
   d. Software learns from conversion
5. Metadata does not conform to LIMBER requirements
   a. Convert to LIMBER version of DDI standard

6. Internationalisation & Localisation

The requirement for the internationalisation of the systems developed in LIMBER requires that they be localised to support an initial set of languages, but more generally that they be internationalised to permit later addition of localisation to other languages. This places a requirement on the development method that it be structured to support the two phases of Internationalisation and Localisation.

6.1. Creating an International User Interface

A major aspect of creating an international user interface involves translating the text used in title bars, menus, other controls, messages, and registry entries. To make this process easier, store interface text as resources in your application's resource file, rather than including it in the source code of the application. Also translate any menu commands that your application stores for its file types in the system registry.
When translating text, remember that each language has its own syntax and grammar. The following are some general guidelines to keep in mind when translating text:

- Avoid using vague words that can have several meanings in different contexts.
- Avoid colloquialisms, jargon, acronyms, and abbreviations.
- Use good grammar. Translation is a difficult task even when a translator does not have to deal with poor grammar.
- Avoid dynamic, or run-time, concatenation of different strings to form new strings—for example, composing messages by combining frequently used strings. An exception is the construction of file names and names of paths.
- Avoid hard-coding file names in a binary file. File names may need to be translated.
- Avoid including text in images and icons. Doing so requires that these also be translated.

Translation of interface text often increases the length of text by 30 percent or more. In some extreme cases, the character count can increase by more than 100 percent; for example, the English word "move" becomes "verschieben" in German. Accordingly, if the amount of space for displaying text is strictly limited, as in a status bar, restrict the length of the interface text to approximately one-half of the available space. In contexts that allow more flexibility, such as dialog boxes and property sheets, allow 30 percent for text expansion in the interface design. Text in message boxes, however, should allow for text expansion of about 100 percent. Avoid having your software rely on the position of text in a control or window because translation may require movement of the text.
Expansion due to translation affects other aspects of your product. A localized version is likely to affect file sizes, which potentially can change the layout of your installation disks and setup software.
Additionally, translation is not always a one-to-one correspondence. A single word can have multiple translations in another language. Adjectives and articles sometimes change their spelling according to the gender of the nouns they modify. Therefore, be careful when reusing a string in multiple places. Similarly, several words may have only a single meaning in another language. This is particularly important when creating keywords for the Help index for your software.

The following Issues should be considered:

addresses and postal codes
calendar representations

character encoding
character fonts
character sets
collating sequence
currency values and format
date formats
gender identification
keyboard keys and layout
measurements
multiple language support
number format
pagination
paper sizes
personal names
punctuation
scan order for page layout
semantics of auditory icons
semantics of color
semantics of visual icons
semantics of visual metaphors
sort order
symbols
time formats
titles
word and sentence delimiters
word order

6.2. Multiple Language Support

The table below lists the top 20 languages in the world, and their usage (in millions of people). To gain maximum coverage of the generic LIMBER tools we could pick the first few languages and localise for those. However, we need to pick the languages which will provide greatest usage to the application of using Social Science Metadata stored in European Archives. Therefore we will pick the top languages in the list where there are archives - English, Spanish French and German. There is no Portuguese social science data archive, so although that language woul improve generic coverage, it would not address the needs of the local social science application. The potential user base for these four languages as official languages is therefore : 2 billion users.

Mother-tongue		Official language		Mother-tongue		Official language	
1. Chinese	1,000	English	1,400	11. French	70	Japanese	120
2. English	350	Chinese	1,000	12. Punjabi	70	German	100
3. Spanish	250	Hindi	700	13. Javanese	65	Urdu	85
4. Hindi	200	Spanish	280	14. Bihari	65	Italian	60
5. Arabic	150	Russian	270	15. Italian	60	Korean	60
6. Bengali	150	French	220	16. Korean	60	Vietnamese	60
7. Russian	150	Arabic	170	17. Telugu	55	Persian	55
8. Portuguese	135	Portuguese	160	18. Tamil	55	Tagalog	50
9. Japanese	120	Malay	160	19. Marathi	50	Thai	50
10. German	100	Bengali	150	20. Vietnamese	50	Turkish	50

As countries that speak other languages high on this list either establish archives, or become potential users of the data in order to make evidence based decisions and plans, then the interfaces should be localised to them. Therefore the internationalisation phase should be able to support such later localisation.

ISO 639 standard two letter codes should be used to represent language names.

6.2.1. Supporting International Characters and Formatting

Character encoding is the most basic foundation for any form of text processing; if it is handled poorly, the software is difficult to localize or internationalise. A program also requires functionality that observes language rules and cultural conventions. Unicode is a universal character encoding system.

6.2.2. Keyboards

When you are internationalizing a user interface, language is not the only factor to consider. Several countries can share a common language but have different conventions for expressing information. In addition, some countries can share a language but use different keyboard conventions.
International keyboards differ. Avoid using punctuation character keys as shortcut keys because they are not always found on international keyboards or easily produced by the user. What seems like an effective shortcut because of its mnemonic association—for example, CTRL+B for Bold—may need to be changed to fit a particular language. Similarly, macros or other utilities that invoke menus or commands based on access keys are not likely to work in an international version because the command names on which the access keys are based differ.
Additionally, keys do not always occupy the same positions on all international keyboards. Even when they do, the interpretation of the unmodified keystroke can be different. For example, on US keyboards, SHIFT+8 results in an asterisk character. However, on French keyboards, it generates the number 8. Similarly, avoid using CTRL+ALT combinations, because the system interprets this combination for some language versions as the ALTGR key, which generates alphanumeric characters. Similarly, avoid using the ALT key as a modifier because it is the primary keyboard interface for accessing menus and controls. In addition, the system uses many specialized versions for special input. For example, ALT+~

invokes special input editors in Asian versions of Windows. For text fields, pressing ALT+number enters characters in the upper range of a character set. Similarly, avoid using the following characters when assigning shortcut keys.
@ £ $ { } [ ] \ ~ | ^ ' < >

6.3. Preparing for Cultural Differences

A more subtle factor to consider when you are preparing software for international markets is cultural differences. For example, users in the US may recognize a rounded mail box with a flag on the side as an icon for a mail program, but this image may not be recognized by users in other countries. Sounds and their associated meanings may also vary from country to country.
It is best to review the proposed graphics for international applicability early in your design cycle. Localizing graphics can be a time-consuming process.
Although graphics communicate more universally than text, graphical aspects of your software—especially icons and toolbar button images—may also need to be revised to address an international audience. For example, a toolbar image that includes a magic wand to represent access to a wizard interface does not have meaning in many countries and requires a different image.
When possible, choose generic images and glyphs. Even if you can create custom designs for each language, having different images for different languages can confuse users who work with more than one language version.
Many symbols with a strong meaning in one culture do not have any meaning in another. For example, many symbols for holidays and seasons are not shared around the world.
Importantly, some symbols can be offensive in some cultures; for example, the open palm commonly used at US crosswalk signals is offensive in some countries. Some metaphors also may not apply in all languages.

6.4. Coding for Internationalisation

When you are coding your application, several coding practices can make the internationalisation process easier. A few of these practices are the following:

- Do not hard-code localizable elements.
  Hard-coded strings, characters, constants, screen positions, file names, and file paths are difficult to track down and localize. Isolate all localizable items into resource files, and minimize compile dependencies.
- Do not make buffers too small to handle localized text.
  Buffers that are declared to be the exact size of a word or a sentence will probably overflow when text is translated. Consider the following example. Your application declares a 2-byte buffer size for the word "OK." In Spanish, however, when it refers to the text in an OK button, the same word is translated as "Aceptar," which would cause your application to overflow.
- Do not perform string composition.
  For example, translating "wrong file" and "wrong directory" to Italian results in "file errat*o*" and "cartella erratt*a*," respectively. If you try to perform string composition using the syntax "wrong%s", it does not work.

Another potential problem involves declaring a single string and displaying it in a number of different contexts: on a menu, in a dialog box, and perhaps in several messages. The problem with using all-purpose strings is that in European languages, adjectives and some nouns have

from 4 to 14 different forms, such as masculine, feminine, and neuter singular; and masculine, feminine, and neuter plural, that must match the nouns they modify. A single string displayed in different contexts is correct in gender and number in some cases but incorrect in others.
One way to ensure that your coding practices works in an international market is to substitute your language strings with a pseudolanguage, and then test your code. Any potential problems should surface immediately.

6.5. Localization Issues for Design and Development

When designing the user interface and when actually implementing the design in the layout and code, there are important issues to consider in order for the pages to be localisable. The following list details some of the more important localisation considerations.

- Add at least 30% more space in the wizard pages and any secondary dialogs for text expansion.
- Add at least one extra line to text boxes when a variable is used within a sentence.
- Try to avoid using UI controls within a sentence.
- Don't use slang and technical jargon.
- Create generic icons and images that don't have to be localized.
- Don't use custom UI controls.
- Make sure all localizable strings are not hard-coded.
- Put font information into the dialog (especially for wizards, the font size at run time is much larger.) Any font information should be in resource (facename, size, charset, codepage langID etc.) when setting font in code.
- Make sure text can be wrapped, especially text after check boxes and radio buttons. This can be accomplished by always setting the text for multi-line and top alignment.
- If possible, don't put the text for buttons etc. in the string table but in the dialogs themselves — not only buttons, but any dialog items and menu items as well.
- Don't combine strings at run time to form new strings.
- Don't separate a sentence string into several strings.
- Get rid of old UI text and controls that are not used anymore.
- Don't use more than one placeholder per message if they are of different types (for example, string and int).
- Enumerate placeholders if more than one is used.
- Don't assume the English structure when the sentence contains placeholders.
- Avoid using placeholders like %s. (If placeholders are used, document what is going to be pasted it).
- Avoid using multiple variables, or use %1, %2, %3… instead of multiple %s.
- Use consistent terminology. Use of period (.), colon (, ; :), and Caps should also be consistent.
- Check spelling (typos in English expressions reduce the chances to use features like auto translate).
- Never place text that must be localized into bitmaps where it cannot be accessed by localization tools.

6.6. ☑ Internationalisation *Check*list

6.6.1. Program specs account for international considerations from the outset.

Features important to international markets are included.
Icons and bitmaps are generic, are culturally acceptable, and do not contain text.
Menu and dialog-box designs leave room for text expansion.
Text and messages are devoid of slang and specific cultural references.
Strings are documented using comments to provide context for translators.
Strings or characters that should not be localized are marked.
Shortcut-key combinations are accessible on international keyboards.
International laws affecting feature designs are considered.
Third-party agreements support international design issues.
Consistent English user interface terminology is used in strings.

6.6.2. Code is generic enough to work for several languages.

Code doesn't concatenate strings to form sentences.
Code doesn't use a given string variable in more than one context.
Code doesn't contain hard-coded character constants, numeric constants, screen positions, filenames, or pathnames that presume a particular language.
Buffers are large enough to handle translated words and phrases.
Program allows input of international data.
All language editions can read one another's documents.
Code contains support for locale-specific hardware, if necessary.
Features that don't apply to international markets can be removed easily.

6.6.3. Code takes advantage of international functionality offered by the Operating System.

Code uses international information carried by the system
Code uses system functions for sorting, character typing, and string mapping.
Code uses generic text layout functions provided by the Multilingual API.
Program responds to changes in Control Panel's international settings.
Far East editions support Input Method Editors, vertical text, and line-breaking rules.

6.6.4. All international editions of the program are compiled from one set of source files.

Mechanisms requiring code to be recompiled for different language editions are weeded out.
Localizable items are stored in resource files.
All language editions using double-byte character sets are based on a single executable.
All language editions using Unicode are based on a single executable.
All bidirectional language editions are based on a single executable..
All language editions share a common file format.

6.6.5. Code is generic enough to handle different character sets

Code properly handles accented characters.
Program handles nonhomogeneous network environments in which machines are running different code pages.
Code for Far East–language applications parses double-byte characters unless the code is based on Unicode.
Code supports Unicode or conversion between Unicode and the local code page.

Code doesn't assume that all characters are 8-bit or 16-bit.
Code uses generic data types and generic function prototypes.
Program displays and prints text using the appropriate fonts.

6.6.6. Program meets international testing standards.

Text is translated and meets the standards of native speakers.
Dialog boxes are resized and text is hyphenated appropriately.
Translated dialog boxes, status bars, toolbars, and menus fit on the screen at different resolutions.
Menu and dialog-box accelerators are unique.
User can type accented characters into documents, dialog boxes, and filenames.
User can type accented characters into documents, dialog boxes, and filenames.
User can successfully cut, paste, save, and print accented characters.
Sorting and case conversion are culturally accurate.
Application works correctly on localized editions of Operating Systems.

7. Bibliography and References

Dennison, K., "Results of Framing the Future II Metadata Tools Workshop"
Miller, K., "Limber WP3 Requirements Workshop Report"

8. Appendix

8.1. UKDA Depositor's Form

Study Description Form

Please read the information given on the previous page before completing this form.

1. Dataset title

Please ensure that the title on the licence form and the title given here are the same.

101

2. Reference number

Any project or reference numbers used by you or your organisation to identify this dataset should be recorded here.

101

☐ Not applicable

3. Principal investigator(s)

Please list the names of the principal investigator(s) or lead researcher(s) or other bodies or persons intellectually responsible for the dataset (use BLOCK CAPITALS). These names will be listed in our catalogue in the order given here.

Forename	Surname	Affiliation (organisation/department) at time of work described by this form.	131

4. Funding

Please provide the names of any person(s) or organisation(s) which funded the creation of the dataset, with grant numbers where appropriate. If funded by a consortium which does not have a distinctive group title, please name all associates.

142

☐ Not applicable

5. Responsibility for data collection

This section should show persons or bodies who hold overall responsibility for the collection or extraction of data especially where this task has been contracted out to specialist data collection agencies. Do not list the names of those who collect data under the supervision of others, or which are the same as, or working under the supervision of those named in Q.3 above.

132

☐ Not applicable

6. Other acknowledgements

Please list the names of any other person(s) or organisation(s) responsible in any significant way for this dataset and indicate their role.

141 99

☐ Not applicable

7. Crown copyright *This section should show details of the holder of the copyright.*

☐ Crown copyright (go to Q.8)

If Crown copyright held jointly with another (please provide details below):

☐ Please provide the name of the holder of the copyright of these data (e.g. an individual person, or the organisation for which they work, or the sponsoring body etc.):

☐ If the copyright is held jointly (eg. by more than one individual or organisation) please give details:

☐Not applicable

8. Abstract

Please provide a brief summary (max. 200-300 words) of the main aims and objectives of the research project from which these data arose. Append on paper or disk if necessary.

201

9.Summary of content

Please provide a brief description (max. 200-300 words) of the content of this dataset.

599

10. Standard measures

If established measurement techniques were employed e.g. the Hope-Goldthorpe occupational desirability scale, or the Likert scales of social and political values, please give details.

599

☐ Not applicable

11. Sources of data

If the data were derived in whole or in part from other published or unpublished sources, whether printed or machine-readable, please give references to the original source material. Please give details of where the sources are held, how they are documented and how they can be accessed. This information may be appended to the form or provided on disk if lengthy.

203

☐ Not applicable (go to Q. 12)

12. Completeness of transcription

Please indicate whether the data represent a complete or partial transcription of the original sources. A photocopy of an example of the original record, with an example of the record as it appears in the database would be helpful.

203

☐ Not applicable

13. Kind of data

Please indicate below the nature of the data materials.

202

- ☐ Textual data
- ☐ Numeric data
- ☐ Alpha/numeric data
- ☐ Image
- ☐ Other, please specify :

Please indicate the level

- ☐ Individual level
- ☐ Aggregate level

14. Temporal characteristics

(a) Please give date(s) of fieldwork or data collection.

231

☐ Not applicable

(b) Please give date(s) when the data were computerised.

300

☐ Not applicable

(c) Please specify time period covered by the data.

220

☐ Not applicable

15. Geographical coverage

Give the name(s) of the country, region, county, town or village covered by the data. Please indicate, for towns or villages, which county or region these came under at the time period covered by the data. Where many areas are covered, please attach a complete list on paper or a disk.

222 23, 35

16. Entities being studied or recorded

Please indicate below the principal units (persons, groups, institutions or other entities) studied or recorded, e.g. primary school teachers, industrial firms, planning applications. Please indicate their defining characteristics, e.g. "Women aged between 18 and 65 working outside the home and earning less than £4 per hour".

222 26, 99

17. Spatial units

If the dataset contains information or variables which will allow analysis by or on spatial units (e.g. wards, parishes), please indicate what these spatial units are. A list of the common spatial units follows. Please tick those which are applicable to your dataset. Extra information may be given in the space provided.

224

- ☐ No spatial units
- ☐ Census tracts
- ☐ Countries
- ☐ County/Scottish Regional Councils
- ☐ Development/planning areas
- ☐ Enumeration Districts
- ☐ Family Health Services Authority Areas
- ☐ Health Authorities
- ☐ Local Authority Areas
- ☐ Local Authority Districts
- ☐ Local Education Authorities
- ☐ Parishes
- ☐ Parliamentary Constituencies
- ☐ Postcode Districts
- ☐ Postcode Sectors
- ☐ Registration Districts
- ☐ Standard Regions
- ☐ TEC/LEC areas
- ☐ Travel-to-work areas
- ☐ Wards
- ☐ Other, please specify :

18. Selection of cases for the dataset

Please describe any data selection or sampling procedures that were applied. If the cases selected were limited by availability, i.e. they were the only cases available, please also state this.

223

☐ No sampling - whole universe covered

Random (Probability) Sampling

☐ Simple random sample
☐ One-stage stratified or systematic random sample
☐ One-stage cluster sample
☐ Multi-stage stratified random sample
☐ Other random (describe) :

Non-random sampling

☐ Quota sample
☐ "Random" walk or other quasi-random sampling
☐ Purposive selection/case studies
☐ Volunteer sample
☐ Convenience sample
☐ Other non-random (describe) :

☐ Not applicable

If the dataset results from a follow-up survey please state selection procedure used for the follow-up sample.

7. Number of cases

If a sample survey, please indicate the selected size of the sample and the achieved number of units. If several samples are included in the dataset, please distinguish between the totals for each, if necessary, on a separate sheet.

Selected sample size : 212

Achieved sample size :

Weighted sample size (if appropriate) :

Weighting factors used (if appropriate) :

20. Method of data collection

Please indicate below the process of data collection. (If computer aided interviewing was used during the data collection process please give details e.g. CATI, CAPI).

232

- ☐ Face-to-face interview
- ☐ Telephone interview
- ☐ Postal survey
- ☐ Other self-completion form
- ☐ Diaries
- ☐ Clinical measurements
- ☐ Physical measurements
- ☐ Psychological measurements
- ☐ Educational measurements
- ☐ Observation
- ☐ Data collected in the course of an administrative activity (see Q. 20)
- ☐ Compilation or transcription of existing material (see Q. 21)
- ☐ Other, please specify :

21. Collection process

If the data were collected in the course of an administrative activity, please provide details below.

232

☐ Not applicable

22. Compilation or transcription of existing material

If the dataset was created using existing data, please provide details below.

232

☐ Not applicable

23. Related datasets

If the dataset is a follow-up survey or is derived from one or more other datasets, please list below the details of the original datasets. If any of these datasets are available from the Archive, it would be helpful if you could indicate the Archive reference number.

411

24. Time dimensions

221

☐ Cross-sectional one-time study

☐ Repeated cross-sectional, please specify how many and how often (actual and intended) :

☐ Follow-up to cross sectional, please specify number of follow-ups :

☐ Panel and cohort survey (including rotational panel), please specify how many waves :

☐ Time series data e.g. result of continuous administrative process or aggregate statistics

☐ Other, please specify :

25. Registration under the Data Protection Act

Please answer the questions below.

Does this dataset contain information which allows identification of the subjects of research ? 111 10

☐ Yes ☐ No

Does this dataset contain information about named individuals, e.g. votes cast for named candidates in elections or the opinions of anonymous survey respondents about named individuals ?

☐ Yes ☐ No

26. Technical reports and documentation

Please list below all technical reports, questionnaires, codebooks and other documentation pertaining to this dataset.

312

27 Resulting Publications

Please provide below full references to any published articles, chapters within books, or complete books, based upon the data in this dataset written by persons or bodies responsible for the dataset. The Archive will be pleased to update this bibliography from time to time, but relies on you to supply the details of any work which is published subsequent to deposit. For this purpose we will send a card for you to complete and return to inform us of further publications. Publications should be quoted in the following style, and can be appended on paper or disk.

Runciman, W.G. (1968) Relative deprivation and social justice London : Routledge & Kegan Paul.

Alt, J., Crewe, I. and Sarlvik, B. (1976) `Partisanship and policy choice : issue preferences in the British electorate, February 1974', British Journal of Political Science Vol. 6 (3), July.

Hill, M.J. (1976) `Can we distinguish voluntary from involuntary unemployment ?' IN G.D.N. Worswick (ed.) The concept and measurement of voluntary employment London : Allen & Unwin.

401

28 Associated publications

Please provide full references to any other publications or reports, based on the data, by persons or bodies other than those originally responsible for the dataset.

411

29. Names of contacts

In case we have difficulty reading or understanding the information you have supplied on this form, we would be grateful for the name, address (in BLOCK CAPITALS) and telephone number of someone who can be contacted for further details.

(1) Name

Address

Tel :
Fax :
email :

(2) Name

Address

Tel :
Fax :
email :

8.2. UKDA Document Types

Good Practice for Documentation Creation

Draft – not to be quoted without permission

The Data Archive, August 1999

Why do we need good documentation?

Introduction.

You may find our recommendations on creating good documentation too stringent, but there are very good reasons why we ask for this level of detail. Good documentation ensures that the research community can use your data to the full; there is less likelihood of misuse or incorrect use; and it may even help you with your original research. It will certainly help you should you wish to return to the dataset for further analysis at some stage in the future.

For primary use:

Consider the benefits at a primary research level:

- Keeping full and detailed records is a matter of good practice and professionalism. Not only will it help others in the future, it will enable you to move on to your next project with a clean slate, knowing that you can return to the data and understand the decisions taken. You may find it hard to remember the name of the project in ten years' time, let alone how you arrived at a particular derived variable.
- Making detailed records of all aspects of the project will ensure that the data are accessible to others working along side you, and will help you to avoid needlessly duplicating another's work.
- Documentation is invaluable in keeping track of changes to data and procedures and helps you identify in retrospect why certain decisions were made.
- Without full and comprehensive documentation, your project is not complete. By recording everything, you can be sure you are offering the absolute best to the research community and you will have a project you can be proud of.
- Increasingly the value of high quality digital resources is being recognised within the community and by the funding councils. A well-presented, well-documented research project will do much to enhance your reputation in the research community.

For secondary use.

And consider the benefits at a secondary research level:

- By providing full and comprehensive documentation you will be offering a valuable insight to secondary researchers on how you conducted your research and arrived at the conclusions you did.
- Good documentation will enable them to replicate your research using the same tools and knowledge that you used.
- It will enable secondary analysts to build upon your research using identical methods.
- By providing good documentation you will be setting standards for others and helping to maintain the quality of research in the community.
- By enabling others to easily access and use your project you will reduce costs by avoiding duplicate data collection efforts.
- As long as they have full documentation, secondary researchers will be able to use your data in new ways in order to produce new research.
- Making fully documented studies available to the research community encourages researchers, from a position of knowledge, to debate the issues your research addresses.
- By providing full methodological details on your research, secondary researchers can test an alternative methodology on the same data.

- A well-documented study is an important resource for students who will benefit from a high-quality, easy-to-use dataset, and may learn valuable insights into good research methodology.
- By keeping standards of documentation high and methodology open to scrutiny the research community will be able to reach consensus on methods.
- Good documentation will ensure that your project remains useable and valuable in perpetuity.

What should be provided?

There are three main types of material that constitute ideal documentation for a dataset:

1. **Explanatory material.** This is the material that is essential to the further, informed use of a dataset. It is the material without which no full understanding of the dataset and its contents can be achieved.
2. **Contextual information.** This provides users with material about the context in which the data were collected and information about the uses to which the data were put. This also forms a vital historical record for researchers of the future. Whilst not essential, inclusion of this information is strongly recommended.
3. **Cataloguing material.** This material serves two purposes. First, it serves as a bibliographic record of the data set. This allows for the dataset to be properly acknowledged and cited in publications arising from an analysis of the dataset, and it acts as a formal record for long preservation purposes. Second, it is the basic instrument used for resource discovery, allowing the dataset to be uniquely identified within the collection by providing appropriate information by which a secondary user can identify the study as useful to their purpose.

Explanatory material:

This section represents the minimum of material that should be created and preserved and can be described as the material required to ensure the long term viability and functionality of a dataset.

1. Information about the data collection methods

This section describes the data collection process, whether a survey, collection of administrative information, transcription from a document source and so on. It should describe the instruments used, the methods employed, and how these were developed. If applicable, details of the sampling design and sampling frames should be included. For example, for a survey, this section would describe the questionnaire, the sampling frame used, the instructions issued to the interviewers, with details of how these were developed and employed during the data collection process. For a collection of administrative data it would include the source of the materials, the transcription process, details of any sampling applied, information on harmonisation or standardisation of the materials and so on. It is also extremely useful to include information on any monitoring process undertaken during the data collection exercise and details of quality controls.

2. Information that describes the structure of the dataset

Key to this type of information is a detailed document that describes the structure of the dataset and that includes information about relationships between individual files or records within the study. It should include, for example, key variables required for unique identification of subjects across files. It should also include the number of cases and variables in each file and the number of files in the dataset. For complex relational models, a diagram showing the structure and the relations between the records and elements of the dataset should be constructed.

Sometimes data creators undertake a re-structuring process on their dataset and in some circumstances the re-structuring may alter the functionality of a dataset. In these circumstances, data creators may wish to consider maintaining both versions with full documentation. Data creators are recommended to discuss which version of the dataset to deposit with the archive and whether it is appropriate to deposit both. Whatever is decided, it is vital that the documentation describes the version(s) deposited.

3. Technical information

This information is basic technical information and should record the computer system on which the files were generated, the software packages with which they have been created, the medium on which the data are stored and a complete list of all data files that make up the dataset. For example, information should record:

- the name and version of a word processing package has been used to create word-processed documents, including any changes and/or conversions that have taken place through the lifecycle of the dataset

- the name and version of the statistical package that has been used to create and to analyse the data, including any changes and/or conversions that have taken place through the lifecycle of the dataset
- the name of the operating system on which the dataset is managed, including any changes or conversions that have occured
- a full record of all the files, together with their sizes

4. Variables and values, coding and classification schemes

It becomes extremely difficult to use a dataset in a sensible and effective way if there is doubt about the meaning of any of the fields or information contained in a dataset. Therefore there should be a complete variable list which describes all the variables (or fields) in the dataset and full explanation, including full details of coding and classifications used, for all the information which can be allocated to those fields. It is especially important to have blank and missing fields explained and accounted for.

Data creators should also identify variables to which standard coding classifications apply and to record the version of the classification scheme used, preferably with a bibliographic reference to that code. Whilst data creators are often very familiar with the appropriate coding classification and have immediate access to it, secondary users may not have this information to hand. Moreover, it may be difficult to remember the version used if returning to the dataset after a period of time.

5. Information about derived variables

Many data producers derive new variables from the original data collected. These may be as simple as grouping raw age data (age in years) to groups of years appropriate for the needs of the survey or something much more complex using a sophisticated algorithm. When grouped or derived variables are created it is important that the logic for the grouping or derivation is clear. Simple grouping, as for age, can be included within the data dictionary but for more complex derivations, other means of recording the information are needed. The best method of describing these is by using flow charts or accurate, Boolean statements. It is recommended that sufficient supporting information be provided to allow an easy link between the core variables used and the resultant variables. We would also recommend that the computer algorithms used to create the derivations be saved together with information on the software used.

5. *Weighting and grossing*

Weighting and grossing variables need to be fully documented, explaining the construction of the variables with a clear indication of the circumstances in which they should be used. The latter is particularly important when different weights need to be applied for different purposes.

6. Data source

Details of the source from which the data are derived should be included in some details. For example, where the data source is responses to survey questionnaires, the text of each question asked should be carefully recorded in the documentation. Ideally the text will include a reference to the variable/s which it has generated. It is also useful to explain the conditions under which a question would be asked of the respondent including, if possible, the number of cases to which it applies and, ideally, summary response statistics. For administrative data, how and why the data were collected should be recorded together with full details of the range and type of information.

7. Confidentiality and anonymisation

It is important to record if the data contain any confidential information concerning individuals, households, organisations or institutions. Where this is the case, then it is recommended that details are noted together with any agreements about how such data might or might not be used, for example, with survey respondents. Issues of confidentiality may restrict what analyses can be undertaken or what can be subsequently published, particularly if the data are to be made available for secondary use.

If the data have been anonymised to prevent identification of subjects, it is wise to record the procedures used and the resultant changes to the data. Such modification may restrict any subsequent analysis and an indication of this is useful.

Contextual information:

This type of information adds richness and depth to the documentation and enables the secondary user to fully understand the background and processes behind the data collection exercise.

1. Description of the originating project
Details should be provided of the history of the project or process giving rise to the dataset. It should provide information on the intellectual and substantive framework of the project giving rise to the data. For example, it might detail why the data collection was felt necessary, the aims and objectives of the project; who or what was being studied; the geographic and temporal coverage; publications arising or policy developments that it contributed to or arose in response to and any other relevant information.

2. Provenance of the dataset
This information is useful in recording the history of the data collection process, changes and developments that occurred, both in the data themselves and the methodology, any adjustments made and so on. It might also record details of data errors, problems encountered in the process of data collection, data entry, data checking and cleaning, conversion to different software or operating system and any other useful information on the life-cycle of the dataset.

Other contextual information is likely to be included in reports or publications that arise from the study and bibliographic references to these should be included.

3. Serial and time-series datasets, new editions
For repeated cross-section, panel or time-series datasets, additional information is required to describe changes across time periods. For survey datasets for example, it is useful to record how frequently a question is asked; whether the question text or response codes have changed; whether a standard classification has been updated during the lifetime of the survey; or whether there has been any change to the composition of the sampling universe of the question. If variables with the same labels represent slightly different information, then this should also be documented. For aggregate data, any changes that make comparison over time unreliable should be recorded.

There are occasions when a dataset is improved or changed and a new version created. For example when harmonising variable names from surveys conducted over several years, the changes need to be documented and the original data maintained so that users can go back to the original data if desired. The reasons for a new version of a dataset need to be recorded and change details maintained.

Finally, where documentation relating to one dataset is useful (or perhaps even essential) for understanding a new, derived or follow-up dataset then a reference to the original documentation should provided together with guidance on which sections of the documentation are applicable.

Cataloguing information:

If datasets are to be made available for other researchers to use, then it is essential that sufficient information is supplied to enable effective resource discovery. Fortunately clear and accepted standards for this type of information exists and it is not too onerous to compile. The information required will include title, principal investigator, sponsors, data collectors, dates of data collection, temporal and geographic coverage, methods of data collection, and sampling design and frames. A summary of the aims and objectives of the data collection project or process and a summary of the content should also be provided. The catalogue record is made available using appropriate search engines, thus promoting the availability of the dataset for secondary analysis. The bibliographic information is used to create a citation statement for the study thereby ensuring that secondary analysts give full recognition to creators of data resources.

Quarterly Labour Force Survey documentation: an exemplar.

The documentation for the Labour Force Survey provides a good example of the type of comprehensive documentation that data producers might provide.

It comprises the following series of volumes:

Volume 1: Background and methodology

This volume contains contextual information about the entire series of Quarterly Labour Force Surveys. It describes the history of the study, the sample and questionnaire design, a detailed account of the fieldwork and the background to the coding and processing of the data. It then goes on to discuss the methodology in detail, and concludes with useful information about publications associated with the series.

Volume 2: Questionnaire

Most of this volume falls into the category of explanatory material and includes the full question text as well as codes indicating where each question was asked, who was asked it and in which quarter. Attached to each question is the variable label and any value labels together with an indication of the routing. Also in this document, associated with each question, are instructions to interviewers which can be categorised as contextual information. Typically, the questionnaire volume will cover four quarters of the LFS.

Volume 3: Details of LFS variables

Similarly, volume 3 gives an account of variables, their labels and routing information. However, this volume goes into much greater detail about who was asked and in which quarter a particular variable was introduced. As this volume covers all years of the study, it is able to show which variables can be tracked over time.

Volume 4: Standard derived variables

Again, a document containing explanatory information, which typically applies to one year's data only. This volume, through the use of flowcharts, shows how derived variables have been constructed.

Volume 5: LFS classifications

Explanatory material, again, is contained in this volume and lists the standard codes needed to make use of the data. It also contains background contextual material on the different groups of standard codes, and reproduces two reports by international statistical bodies and other recognised sources on the various issues surrounding international codes.

Volume 6: Local area data user guide

This volume is the user guide for an aggregated set of data drawn from the main data. It contains both contextual and explanatory material which includes a background to the LAD and how it relates to the main data. It also lists variables and coding specific to the LAD.

Volume 7: LFS variables 1979-1991

Again, this volume gives a small amount of background contextual information and then lists all variables over time in two batches, 1979-1983 (biennial studies) and 1984-1991 (annual), which cover the years of the LFS before it became quarterly in March 1992. This volume makes it possible to compare variables in the earlier years of the LFS with those produced in the quarterly datasets.

Volume 8: Household and family data user guide

A volume containing both contextual and explanatory material on the household and family data which is a set of data, at household level, which has been derived from the main quarterly data. It gives detailed instructions on how to create, combine and analyse several variables using either Quanvert or SPSS.

Volume 9: Eurostat and Eurostat derived variables

This volume contains contextual material describing Eurostat and how it relates to the UK Labour Force Survey. It also contains detailed explanatory material including derived variable flowcharts and a list of codes and questions used.

8.3. UKDA DDI Metadata Catalogue Record


```
<!-- Formatted: 05/01/2000 14:35 GMT -->
<codeBook>
<docDscr>
<citation>
<titlStmt>
<titl>XML codeBook for SN:67017</titl>
</titlStmt>
<prodStmt>
<prodDate date="1999-02-02">2 February 1999</prodDate>
</prodStmt>
</citation>
</docDscr>
<stdyDscr>
<citation>
<titlStmt>
<titl>Images of the World in the Year 2000</titl>
<subTitl>Great Britain National</subTitl>
<IDNo agency="UKDA">67017</IDNo>
</titlStmt>
<rspStmt>
<AuthEnty affiliation="University of Essex. Department of
Sociology">Matthews, D.</AuthEnty>
<AuthEnty>Jenkins, R.</AuthEnty>
<dataCollector>Research Services Limited</dataCollector>
</rspStmt>
<prodStmt>
</prodStmt>
<distStmt>
<dataDist abbr="UKDA" affiliation="University of Essex, Wivenhoe Park,
Colchester, Essex, England, CO4 3SQ">The Data Archive</dataDist>
<depositr affiliation="University of Essex. Department of
Sociology">Matthews, D.</depositr>
</distStmt>
<serStmt>
</serStmt>
</citation>
<stdyInfo>
<subject>
<keyword>AGE factual</keyword>
<keyword>AGE DIFFERENCES attitudinal</keyword>
<keyword>AGGRESSIVENESS attitudinal</keyword>
<keyword>ALLIANCES attitudinal</keyword>
<keyword>ARMED FORCES attitudinal</keyword>
<keyword>BIRTH CONTROL attitudinal</keyword>
<keyword>BRITISH POLITICAL PARTIES attitudinal</keyword>
<keyword>CANCER attitudinal</keyword>
<keyword>CAPITALIST SYSTEMS attitudinal</keyword>
<keyword>CAREER DEVELOPMENT attitudinal</keyword>
<keyword>COLLECTIVIST ECONOMY attitudinal</keyword>
<keyword>COLONIALISM attitudinal</keyword>
<keyword>CURRENCY attitudinal</keyword>
<keyword>DEMOCRACY attitudinal</keyword>
<keyword>DEVELOPING COUNTRIES attitudinal</keyword>
<keyword>DISARMAMENT attitudinal</keyword>
<keyword>ECONOMIC AND SOCIAL DEVELOPMENT attitudinal</keyword>
<keyword>ECONOMIC SYSTEMS attitudinal</keyword>
<keyword>EDUCATIONAL BACKGROUND factual</keyword>
<keyword>EUROPEAN ECONOMIC COMMUNITY attitudinal</keyword>
```

```
<keyword>EUROPEAN UNION attitudinal</keyword>
<keyword>FAMILY SIZE attitudinal</keyword>
<keyword>FATHERS factual</keyword>
<keyword>FORECASTING attitudinal</keyword>
<keyword>FOREIGN POLICY attitudinal</keyword>
<keyword>FUTURE attitudinal</keyword>
<keyword>GENDER attitudinal</keyword>
<keyword>GENDER factual</keyword>
<keyword>GOVERNMENT attitudinal</keyword>
<keyword>HOUSEHOLDS factual</keyword>
<keyword>HUNGER attitudinal</keyword>
<keyword>INCOME attitudinal</keyword>
<keyword>INCOME factual</keyword>
<keyword>INDUSTRIES attitudinal</keyword>
<keyword>INTERETHNIC RELATIONS attitudinal</keyword>
<keyword>INTERNAL POLITICS attitudinal</keyword>
<keyword>INTERNATIONAL CONFLICT attitudinal</keyword>
<keyword>INTERNATIONAL COOPERATION attitudinal</keyword>
<keyword>INTERNATIONAL EQUILIBRIUM attitudinal</keyword>
<keyword>INTERNATIONAL LANGUAGES attitudinal</keyword>
<keyword>INTERNATIONAL ORGANIZATIONS attitudinal</keyword>
<keyword>INTERNATIONAL RELATIONS attitudinal</keyword>
<keyword>INTERNATIONAL TENSION attitudinal</keyword>
<keyword>INTERPERSONAL RELATIONS attitudinal</keyword>
<keyword>ISOLATIONISM attitudinal</keyword>
<keyword>JOB SATISFACTION attitudinal</keyword>
<keyword>MARITAL STATUS factual</keyword>
<keyword>MEDICAL TREATMENT attitudinal</keyword>
<keyword>MEMBERSHIP factual</keyword>
<keyword>MILITARY SERVICE factual</keyword>
<keyword>MINISTERS attitudinal</keyword>
<keyword>MIXED ECONOMY attitudinal</keyword>
<keyword>NATO attitudinal</keyword>
<keyword>NEWSPAPERS attitudinal</keyword>
<keyword>NUCLEAR WEAPONS attitudinal</keyword>
<keyword>OCCUPATIONS factual</keyword>
<keyword>ORGANIZATIONS factual</keyword>
<keyword>PEACE attitudinal</keyword>
<keyword>PEACE-KEEPING FORCES attitudinal</keyword>
<keyword>PEACEFUL COEXISTENCE attitudinal</keyword>
<keyword>PERSONALITY attitudinal</keyword>
<keyword>PERSONALITY factual</keyword>
<keyword>POLITICAL ALLEGIANCE attitudinal</keyword>
<keyword>POLITICAL ATTITUDE attitudinal</keyword>
<keyword>POLITICAL SYSTEMS attitudinal</keyword>
<keyword>POPULATION factual</keyword>
<keyword>POPULATION MIGRATION attitudinal</keyword>
<keyword>POVERTY attitudinal</keyword>
<keyword>PRIVATE OWNERSHIP attitudinal</keyword>
<keyword>PUBLIC OPINION attitudinal</keyword>
<keyword>PUBLIC OWNERSHIP attitudinal</keyword>
<keyword>RELIGION attitudinal</keyword>
<keyword>RELIGIOUS BELIEF factual</keyword>
<keyword>SATISFACTION attitudinal</keyword>
<keyword>SCIENTIFIC PROGRESS attitudinal</keyword>
<keyword>SIBLINGS factual</keyword>
<keyword>SOCIAL CLASS factual</keyword>
<keyword>SOCIAL ISOLATION attitudinal</keyword>
<keyword>SPACE EXPLORATION attitudinal</keyword>
<keyword>TECHNICAL ASSISTANCE attitudinal</keyword>
<keyword>WAR attitudinal</keyword>
```

```
<keyword>WAR CASUALTIES factual</keyword>
<keyword>WAR DISADVANTAGED factual</keyword>
<keyword>WARSAW PACT attitudinal</keyword>
<keyword>WEALTH attitudinal</keyword>
<keyword>WEATHER MODIFICATION attitudinal</keyword>
<keyword>WORKING MOTHERS factual</keyword>
<keyword>WORLD GOVERNMENT attitudinal</keyword>
<topcClas>XXVI\A\2\(a) - General studies - Non-UK countries - Public
opinion - Political behaviour and attitudes</topcClas>
<topcClas>XXVI\A\1\(a) - General studies - United Kingdom - Public opinion
- Political behaviour and attitudes</topcClas>
</subject>
<abstract>This inquiry into the views of the year 2000 held by the younger
generation took place under the auspices of the European Coordination
Centre for Research and Documentation in the Social Sciences, established
at Vienna, which was founded by UNESCO and which is a division of the
International Social Science Council at Paris. The technical coordination
was in the hands of the International Peace Research Institute, Oslo,
under the direction of Johan Galtung.</abstract>
<abstract>To examine attitudes of people in the age group 15 - 40 years
towards various aspects of the future, with particular reference to war,
peace and disarmament. The great attractiveness of such an inquiry lies in
comparing the results of countries with very different political and
philosophical backgrounds. Eleven countries are covered by this
study.</abstract>
<abstract>Attitudinal/Behavioural Questions</abstract>
<abstract>Respondent's future-consciousness is assessed in terms of his
thinking about the future of the world and of his country, his perception
of the year 2000 as the near or distant future, his talking, seeing,
hearing and reading about the future. Respondent predictions: what he
considers will be the main differences between life today and life in the
year 2000 (particularly what he feels would be the best and worst things
that could happen). Employing a 9-point scale (i.e. 'best' - 'worst'
possible life) the respondent is requested to indicate where he would
place himself: a) at the present time, b) five years ago, c) five years
from now, d) in the year 2000. Using the same procedure he is asked to
assess future trends of his country and of the world.</abstract>
<abstract>More specifically, the respondent is asked to predict social
trends in his country covering topics such as: happiness and work
satisfaction, leisure, unemployment, religion, kinship and marriage,
material wealth, spiritual contentment, sexual freedom, mental illness,
use of narcotics and drugs, crime, social differentiation, the role of
women, the role of young people, city dwelling and manual work. It is
recorded whether, in most cases, the respondent's hopes coincide with his
predictions.</abstract>
<abstract>Respondent predictions of the possibilities of science in the
year 2000 are ascertained. Namely, whether it will be possible: to
predetermine the sex and major personality feature of one's child, to cure
dangerous diseases (e.g. cancer), to predetermine the weather, to travel
to other planets. The respondent is again asked to state whether his hopes
coincide with his predictions.</abstract>
<abstract>War, armament and disarmament: respondent assessments of world
trends in this area are recorded. In addition, he is asked to assess the
probable effects of a third world war on his native country, and to state
his opinion on how such a war is most likely to break out (i.e. by
accident, by extension of a limited conflict or by one big power attacking
another big power). Any value, goal or ideal the respondent believes could
justify a war with nuclear weapons/without nuclear weapons is noted. A
list of 25 ideas on how world peace might be obtained is included and
respondents are asked to state whether they agree or disagree with each
statement (e.g. 'to obtain peace, hunger and poverty must be abolished all
```

over the world', 'to obtain peace, we must have general and complete disarmament as soon as possible', etc.). Information also includes whether the respondent thinks that peace can be realised by the year 2000 and whether he believes he can contribute anything himself to the realisation of this proposal; what he believes is most likely to happen in the relations between capitalist and socialist countries, between rich and poor countries and between different races. Finally, respondent's knowledge of the membership of NATO and the Warsaw Pact is tested.</abstract>
<abstract>Opinion is ascertained on a number of items tapping the personality of the respondent (e.g. dogmatism). Social satisfaction of the respondent is measured in regard to income, job, influence on public affairs, living in his country, whether the respondent believes he has control over his future and, if so, how he feels he should direct this future. He is also asked to comparatively evaluate certain activities and views of the younger and older generations.</abstract>
<abstract>Background Variables</abstract>
<abstract>Age, sex, marital status, education, occupational details, work satisfaction (ideal occupation is noted), personal monthly income quartile, satisfaction with income received, occupation of head of household (where different), total monthly income quartile of household, household composition, area of residence (i.e. density of population, geographical region - where available), whether respondent practises religion or considers himself to be a `believer', parental household composition, father's occupational details, whether mother worked outside the home, area of childhood residence (i.e. density of population, geographical region - where available), age at which respondent moved away from parental home, and finally, details of the respondent's organisational membership is given.</abstract>
<abstract>Attitudinal/Behavioural Questions</abstract>
<abstract>The section in the standard questionnaire (para 2) on predictions of social change in respondent's native country is excluded in the British version, the emphasis being more particularly on international relations and politics.</abstract>
<abstract>The following is added:</abstract>
<abstract>Aspects of British foreign policy are considered, for example, whether she should join the Common Market and if she does, whether she should retain the right to decide her own internal affairs. Respondent opinion is also ascertained on the general type of foreign policy he would like to see Britain pursue (e.g. whether there are any countries the respondent feels Britain should have no contact with). Factors influencing the formation of foreign policy are considered together with the role of the United Nations (particularly whether the UN should have the power to intervene between Britain and some other country, whether the UN should have the power to intervene in British internal affairs).</abstract>
<abstract>Disarmament: views on disarmament and how this should be achieved are recorded, (i.e. whether all weapons should be abolished at the same time, whether disarmament is easy/hard to achieve etc.).</abstract>
<abstract>Military alliances and political tension: the country, or group of countries, the respondent feels is mainly responsible for the political tension in the world today is noted, together with whether he sees the military forces of the Soviet Union and the socialist powers as a threat, and whether he thinks that NATO is seen by other countries as a threat. He is also asked to state which of the Eastern and Western military powers he considers to be the strongest in Europe, to list the factors that he considers to have been of importance in bringing about a relaxation of tension in Europe, and to generally assess the trends in the relations between Eastern and Western European nations.</abstract>
<abstract>The following data are recorded for the respondent's childhood home: place of residence, household composition (at the age of 14 years),

father's occupational details, whether mother worked outside the home, age at which respondent moved away from parental home, number of times parents moved district (up until the age of 14 years), political interest in the home (and whether respondent now tends to agree or disagree with his parents' political views).</abstract>
<abstract>Background Variables</abstract>
<abstract>Age, sex, marital status, social grade, number of children respondent expects to have, educational achievements, number of years full-time education received, occupational details (including the number of years in present occupation, job satisfaction, respondent's ideal job in the year 2000), occupational details of the head of household (if different), personal monthly income, household monthly income, place of residence, length of residence, religious belief, political and religious organisational membership, whether respondent was required to do military service, whether he, or any member of his family, was actively involved in or directly affected by, the second world war, finally, the respondent's voting intention, were there a general election tomorrow, is recorded, and whether he has ever voted differently.</abstract>
<sumDscr>
<timePrd date="1967-00-00" event="single">1967</timePrd>
<collDate date="1967-08-00" event="start">August 1967 </collDate>
<collDate date="1967-09-00" event="end">September 1967 </collDate>
<nation>Cross-national</nation>
<nation>Great Britain national</nation>
<geogCover>GREAT BRITAIN</geogCover>
<anlyUnit>Individuals</anlyUnit>
<universe level="study">Adults</universe>
<universe level="study">Cross-national</universe>
<universe level="study">National</universe>
<universe level="study">British population 15 - 40 years old</universe>
</sumDscr>
</stdyInfo>
<method>
<dataColl>
<timeMeth>Cross-sectional (one-time) study</timeMeth>
<sampProc>Two stage sample - first, probability sample of local authority administrative areas, the second stage a quota sample (sex, age, social grade) </sampProc>
<deviat>1000 (target) 1001 (obtained) </deviat>
<collMode>Face-to-face interview</collMode>
<sources>
</sources>
</dataColl>
</method>
<dataAccs>
<setAvail>
<accsPlac>Data Archive</accsPlac>
</setAvail>
<useStmt>
<restrctn>Depositor has specified :- Special access conditions apply. Details available from the Archive and normally involve signing an undertaking form specific to this study.; Because these data contain details of identifiable individuals or organisations, users should register their research with the Data Protection Registrar</restrctn>
</useStmt>
</dataAccs>
<othrStdyMat>
<relStdy>Group constituents: 1226, 67017-67025, 68009-68010, 69019</relStdy>
<relStdy>Group: 33009</relStdy>

```
<relPubl>Ornauer, H., et al, "Images of the world in the year 2000" (The
Hague: Mouton, 1976; New York: Humanities Press) </relPubl>
<othRefs>Study description: English; Research instrument: English;
Codebook: English</othRefs>
<othRefs>Number of Cases: 1001 cases   </othRefs>
</othrStdyMat>
</stdyDscr>
<fileDscr>
<fileTxt>
<fileName>d67017.asc</fileName>
<fileStrc type="rectangular"/>
<dimensns>
<caseQnty>0</caseQnty>
<varQnty>214</varQnty>
<recPrCas>1</recPrCas>
</dimensns>
<fileType>ASCII Tab Delimited</fileType>
</fileTxt>
</fileDscr>
<dataDscr>
<var name="RESP" intrvl="discrete" format="int" dcml="0">
<location StartPos="1"/>
<labl level="variable">RESPONDENT NUMBER</labl>
</var>
<var name="NI" intrvl="discrete" format="int" dcml="0">
<location StartPos="2"/>
<labl level="variable">NATIONAL IDENTIFIER</labl>
</var>
<var name="CARDNO" intrvl="discrete" format="int" dcml="0">
<location StartPos="3"/>
<labl level="variable">DECK NUMBER</labl>
</var>
<var name="Q1" format="char" dcml="0">
<location StartPos="4"/>
<labl level="variable">THINK ABOUT FUTURE OF COUNTRY IN YR 2000?</labl>
<qstn><qstnLit>HOW MUCH WOULD YOU SAY THAT YOU THINK ABOUT THE FUTURE OF
YOUR COUNTRY,
NOT IN A COUPLE OF YEARS BUT, SAY, IN THE YEAR 2000?
</qstnLit></qstn>
<catgryGrp>
<catgry>
<catValu>1</catValu>
<labl level="category">Very much</labl>
</catgry>
<catgry>
<catValu>2</catValu>
<labl level="category">Some</labl>
</catgry>
<catgry>
<catValu>3</catValu>
<labl level="category">A little</labl>
</catgry>
<catgry>
<catValu>4</catValu>
<labl level="category">Not at all</labl>
</catgry>
<catgry>
<catValu>9</catValu>
<labl level="category">Don&apos;t know</labl>
</catgry>
</catgryGrp>
```

```
</var>
<var name="Q2" format="char" dcml="0">
<location StartPos="5"/>
<labl level="variable">THINK ABOUT FUTURE OF WORLD IN YR 2000?</labl>
<qstn><qstnLit>HOW MUCH WOULD YOU SAY THAT YOU THINK ABOUT THE FUTURE OF
THE WHOLE
WORLD, NOT IN A COUPLE OF YEARS, BUT, SAY, IN THE YEAR 2000?
</qstnLit></qstn>
<catgryGrp>
<catgry>
<catValu>1</catValu>
<labl level="category">Very much</labl>
</catgry>
<catgry>
<catValu>2</catValu>
<labl level="category">Some</labl>
</catgry>
<catgry>
<catValu>3</catValu>
<labl level="category">A little</labl>
</catgry>
<catgry>
<catValu>4</catValu>
<labl level="category">Not at all</labl>
</catgry>
<catgry>
<catValu>9</catValu>
<labl level="category">Don&apos;t know</labl>
</catgry>
</catgryGrp>
</var>
<var name="Q3" format="char" dcml="0">
<location StartPos="6"/>
<labl level="variable">FEEL YR 2000 FAR AWAY IN DISTANT FUTURE?</labl>
<qstn><qstnLit>SOME PEOPLE FEEL THAT THE YEAR 2000 IS FAR AWAY IN THE
DISTANT FUTURE,
OTHERS FEEL IT IS RATHER CLOSE IN THE NEAR FUTURE. WHAT DO YOU FEEL
ABOUT THIS?
</qstnLit></qstn>
<catgryGrp>
<catgry>
<catValu>*</catValu>
<labl level="category">Invalid punch</labl>
</catgry>
<catgry>
<catValu>1</catValu>
<labl level="category">Distant</labl>
</catgry>
<catgry>
<catValu>2</catValu>
<labl level="category">Uncertain/Don&apos;t know</labl>
</catgry>
<catgry>
<catValu>3</catValu>
<labl level="category">Close</labl>
</catgry>
</catgryGrp>
</var>
<var name="Q4C1" format="char" dcml="0">
<location StartPos="7"/>
<labl level="variable">OFT TALK ABOUT FUTURE OF COUNTRY/WORLD?</labl>
```

```
<qstn><qstnLit>HOW OFTEN WOULD YOU SAY THAT YOU:
TALK WITH SOMEBODY ABOUT THE FUTURE OF YOUR COUNTRY OR THE WORLD?
</qstnLit></qstn>
<catgryGrp>
<catgry>
<catValu>*</catValu>
<labl level="category">Invalid punch</labl>
</catgry>
<catgry>
<catValu>1</catValu>
<labl level="category">Never</labl>
</catgry>
<catgry>
<catValu>2</catValu>
<labl level="category">Less than once a mth</labl>
</catgry>
<catgry>
<catValu>3</catValu>
<labl level="category">Once a month</labl>
</catgry>
<catgry>
<catValu>4</catValu>
<labl level="category">Once a week</labl>
</catgry>
<catgry>
<catValu>5</catValu>
<labl level="category">More often</labl>
</catgry>
<catgry>
<catValu>9</catValu>
<labl level="category">Don&apos;t know</labl>
</catgry>
</catgryGrp>
</var>
<var name="Q4C2" format="char" dcml="0">
<location StartPos="8"/>
<labl level="variable">OFT WATCH/LISTEN TO ITEMS ABOUT FUTURE?</labl>
<qstn><qstnLit>HOW OFTEN WOULD YOU SAY THAT YOU:
WATCH OR LISTEN TO ITEMS ABOUT THE FUTURE OF YOUR COUNTRY OR THE
WORLD IN RADIO OR TV?
</qstnLit></qstn>
<catgryGrp>
<catgry>
<catValu>*</catValu>
<labl level="category">Invalid punch</labl>
</catgry>
<catgry>
<catValu>1</catValu>
<labl level="category">Never</labl>
</catgry>
<catgry>
<catValu>2</catValu>
<labl level="category">Less than once a mth</labl>
</catgry>
<catgry>
<catValu>3</catValu>
<labl level="category">Once a month</labl>
</catgry>
<catgry>
<catValu>4</catValu>
<labl level="category">Once a week</labl>
```

```
</catgry>
<catgry>
<catValu>5</catValu>
<labl level="category">More often</labl>
</catgry>
<catgry>
<catValu>9</catValu>
<labl level="category">Don&apos;t know</labl>
</catgry>
</catgryGrp>
</var>
<var name="Q4C3" format="char" dcml="0">
<location StartPos="9"/>
<labl level="variable">HOW OFT READ ABOUT FUT IN NEWSPAPER/BOOK?</labl>
<qstn><qstnLit>HOW OFTEN WOULD YOU SAY THAT YOU:
READ ABOUT THE FUTURE OF YOUR COUNTRY OR THE WORLD IN A NEWSPAPER
OR A BOOK?
</qstnLit></qstn>
<catgryGrp>
<catgry>
<catValu>*</catValu>
<labl level="category">Invalid punch</labl>
</catgry>
<catgry>
<catValu>1</catValu>
<labl level="category">Never</labl>
</catgry>
<catgry>
<catValu>2</catValu>
<labl level="category">Less than once a mth</labl>
</catgry>
<catgry>
<catValu>3</catValu>
<labl level="category">Once a month</labl>
</catgry>
<catgry>
<catValu>4</catValu>
<labl level="category">Once a week</labl>
</catgry>
<catgry>
<catValu>5</catValu>
<labl level="category">More often</labl>
</catgry>
<catgry>
<catValu>9</catValu>
<labl level="category">Don&apos;t know</labl>
</catgry>
</catgryGrp>
</var>
</dataDscr>
</codeBook>
```

8.4. Thesaurus Management System Evaluation Criteria

by Jochen Ganzmann

Originally published in ***International Classification*, 1990, vol.17, no.3/4, p.155-157** as an appendix to the paper

A. GENERAL CRITERIA

1. Technical Data

- Hardware Compatibility:
  - computers on which software runs
  - storage required:
    - RAM
    - external storage devices
  - operating systems
    - single user
    - multi-user
- Software Package
  - programming language
  - single user
  - multi-user

2. Development Data

- Developer
- Versions:
  - recent version
  - first version
  - overall number of versions

3. Prices

- Software Package:
  - single user
  - multi-user
- Extras/Modifications
- Updates:
  - single user
  - multi-user
- Support:
  - installation
  - updating
  - application
  - hotline
- Training
- Discounts

4. Support

- Supporting Institution
- Forms of support
  - hotline telephone
  - consultation
  - training
  - newsletters
  - active support
    - installation
    - updating
    - modification

5. Acceptance

- Number of installations
- User groups
- Reviews in articles

6. Ergonomics

6.1 Documentation

- Types of manual:
  - operations manual
  - user manual
- Parts included:
  - table of contents
  - documentation of:
    - technical specifications
    - installation
    - application
    - error messages
    - backup and recovery
  - index
- User friendliness
  - structure of manual
  - completeness of information
  - correctness of information
  - clarity:
    - style
    - examples
  - training disc
  - tutorial

6.2 Software Ergonomics

- Language of User Surface
- Complexity of Screen Layout:
  - structure of information
  - colouring
  - window technique
- Dialog Forms:
  - command driven
  - menu driven
  - hybrid
  - mouse
- Help Functions
- Messages:
  - self-explanatory
  - explained in manual
  - error messages
  - feedback messages
  - alert messages
  - confirmation messages
- Provision for Different User Levels

7. Data Integrity

- Access Control:
  - password
  - restrictions for individual users
    - restriction to specific databases
    - restrictions to specific functions
- Backup Procedures
  - automatic

    - storage device
- Reorganisation Features
- Recovery Features

B. CRITERIA RELATING TO FUNCTIONS OF THESAURUS SOFTWARE

1. Structural Definitions

1.1 Term and Term Related Attributes

- Predefined Fields for:
    - Term
        - maximum number of characters
    - Scope Note/Text
        - maximum number of characters
    - Notation
        - no differentiation
        - differentiation for:
            - broad categorization (subject groups/facets)
            - systematic categorization
        - maximum number of characters
    - Source of Term
        - maximum number of characters
        - variable length
    - Information as to Language of Term
        - maximum number of characters
    - additional fields
        - maximum number of characters
- User Definitions
    - number of fields
    - length of fields
    - sequence of fields

1.2 Relations

1.2.1 Among Terms of One Vocabulary (Monolingual Thesaurus)

- Definition of Relations:
    - predefined relations
    - relations user-definable
- Number of Predefined relations
- Types of Relations:
    - equivalence relationship:
        - normal synonymy (non-descriptor(s) → descriptor)
        - semantic factoring (non-descriptor → descriptors)
        - alternatives (non-descriptor → alternative descriptors)
    - hierarchical relationship:
        - no differentiation
        - differentiation of partitive and generic relation
        - definition of dividing principles (categories)
    - associative relationship:
        - no differentiation
        - differentiation of various types (e.g. predecessor - successor, appurtenance relation etc.)
            - which relations?
- Number of Relations between Individual Terms:
    - equivalence relationship:
        - normal synonymy (max. number of non-descriptors per descriptor)
        - semantic factoring (max. number of factors per non-descriptor)
        - alternatives (max. number of alternative descriptors per non-descriptor)

- hierarchical relationship:
  - number of lower terms per broader term
  - number of broader terms per lower term (polyhierarchy)
  - number of hierarchical levels
- associative relationship

1.2.2 Among Terms from Different Vocabularies

- Type of Vocabularies:
  - multilingual thesauri
  - compatible vocabularies
- Connection Between Different Natural Languages (Multilingual Thesauri)
  - maximum number of different languages
  - status of individual language(s):
    - equal languages
    - dominance of one language
- Connection between Different Indexing Languages:
  - maximum number of indexing languages
  - types of indexing language:
    - classifications
    - thesauri
  - status of individual language
- Mode of Connection:
  - reference of terms to a switching language
  - direct translation of different vocabularies (mapping of vocabularies)

2. Input (Thesaurus Construction and Maintenance)

2.1 Capture of Data

- Mode of Capture:
  - batch input from other system
  - keyboard:
    - mode of input of terms and attributes
    - mode of input of relations
- Ease of Capture:
  - complexity of input of terms and relations
    - separate steps?
    - fixed sequence of input routines?
  - display of entered terms (and relations) on screen
  - automatic derivation of implicit relations

2.2 Modification

- Mode of Modification:
  - global changes possible (of language codes etc.)
  - keyboard
    - mode of modification of terms and attributes
    - mode of modification of relations
- Ease of Modification:
  - complexity of modification
  - ease of changes affecting the status of terms (descriptor - non-descriptor)
  - display of terms (and relations) on screen

2.3 Deletion

- Mode of Deletion:
  - global deletions of terms/relations
  - keyboard
    - mode of deletion of terms and attributes
    - mode of deletion of relations
- Ease of Deletion
  - complexity of deletion
  - automatic deletion of relations of a term deleted

2.4 Consistency Controls

- Definition:
  - predefined
  - user-definable
- Term and Term Attributes:
  - rejection of duplicate entries of the same term
    - modification of control possible for input of several natural or indexing languages
  - definition of admissibility of characters for attribute fields (language codes, notation etc.)
- Relations:
  - control of reciprocity of relations
  - rejection of more than one type of relation between two terms
  - rejection of incomplete relations (e.g. semantic factoring with only one factor)
  - rejection of duplicate relations of one type between two terms
  - rejection of hierarchical or associative relationship between descriptors and non-descriptors
  - control of illogical relations across hierarchical levels
  - other controls

3. Output

3.1 Display on the Screen

- Mode of Search for Terms:
  - browsing
  - scrolling
  - other possibilities
- Display of Individual Terms
  - with attributes
  - with relations
- Display of Word-Lists
  - criteria for selection of terms:
    - alphabetical section
    - strings
    - attributes (language, notation, source etc.)
    - types of relation
    - words marked for specific purposes
    - combination of criteria
  - forms of display of word-lists:
  - alphabetical array:
    - word-list
    - word-list plus relations and attributes
    - other variations
  - KWIC-display
  - hierarchical display
  - systematic presentation (sorting by notation)
    - detailed system
      - without reference to relations
      - with reference to relations
    - broad categories (subject groups/facets)
  - graphical display
- Interaction Possible in Thesaurus on Screen:
  - scrolling/browsing
  - navigation to semantically related terms
  - selection of terms for editing and deletion
  - direct modifications and deletions in lists

3.2 Output by the Printer

- Definition of Output Formats:
  - standard formats predefined

- user definable formats
  - storage of user defined formats
- Criteria for Selection of Terms:
  - alphabetical section
  - strings
  - attributes (notation, facet etc.)
  - types of relation
  - combination of criteria
- Forms of Display:
  - alphabetical array
    - without further information
    - with relations
    - with attributes
  - KWIC-display
  - hierarchical display
    - without relations
    - with relations
  - systematic presentation (sorting by notation)
    - detailed system
      - without relations
      - with relations
      - with attributes
      - with node labels
    - broad categories (subject groups/facets)
  - graphical display
  - display in columns for multilingual/compatible vocabularies
- User-definable Features:
  - information added to terms:
    - relations
    - attributes
  - presentation of the relations:
    - suppression of certain relations (e.g. implicit relations)
    - sequence of relations in print
    - user-definable reference codes for output (e.g in accordance with ISO/DIN)
  - layout:
    - pagination
    - line pitch
    - caption
    - typographic differentiation of descriptors/non-descriptors
    - other features

3.3 Output to a File

- Formats of Output:
  - ASCII file
  - Special format required by other system (i.e. retrieval software, thesaurus maintenance program)

4. Indexing and Retrieval

4.1 Indexing

- Orientation:
  - display forms of thesaurus on screen (cf. also 3.1: Display on the screen):
    - alphabetical display
    - systematic display
    - other forms of display
  - search mode for terms
  - navigation through semantic structure
- Mode of input:

    - entering of terms
    - direct selection of terms from screen thesaurus
- Control of Input:
    - rejection of unknown terms
    - user-definable for use of candidate terms
- replacement of thesaurus terms not admitted for indexing:
    - replacement of compound terms by semantic factors
    - replacement of non-descriptor by descriptor (for thesauri with preferred terms)
    - replacement of terms in secondary language by terms from dominant language in multilingual thesauri
- Representation of concepts:
    - preferred term (descriptor)
    - no preferred term
- Updating:
    - global changes in index
    - statistics on use of descriptors

4.2 Retrieval

- Orientation:
    - display forms of thesaurus on screen (cf. also 3.1: Display on the screen):
        - alphabetical display
        - systematic display
        - other forms of display
    - search mode for terms
    - navigation through semantic structure
- Mode of input:
    - entering of terms
    - direct selection of terms from screen thesaurus
- Control of input:
    - rejection of unknown terms
    - replacement of thesaurus terms not admitted for the representation of concepts:
        - replacement of compound terms by semantic factors
        - replacement of non-descriptors by descriptors (in thesauri with preferred terms)
        - replacement of terms from secondary language by terms from dominant language in multilingual thesauri
    - automatic inclusion of all synonyms (in case of thesauri without preferred terms)
- Formulation of search strategies:
    - automatic generic search option
    - automatic search for related terms
    - automatic inclusion of search term predecessors
- Updating:
    - statistics on the use of search terms

8.5. Table of performance of current commercially available TMS against evaluation criteria.

The following table lists some of the features which ought to be considered when choosing a software package for thesaurus development. It should be read in conjunction with the descriptive information about each package and the general notes given there.
This table includes only those packages which can be bought as stand-alone software, not part of a complete database management system. If a thesaurus is being developed for use with a particular information storage and retrieval system, it is important that the combined system should be evaluated as a whole.
In most cases objective data have been given in this table; these have been obtained either from the suppliers' documentation of by experiment. For some of the ergonomic aspects a subjective rating for quality has been given as a mark out of 5, shown in square brackets, e.g. "[3]".

Name of package	**Beat**	**MultiTes**	**STRIDE**	**TAT**	**TCS**	**Term Manager**	**Hierarch**
Terms and their attributes							
Maximum term length	60	125	80	120	60	63	255
Length of scope notes	2000	64000	1024	32000	64000	630	unlimited
Classification codes	yes	yes	can be implemented as a user-defined relationship	no	no	yes	yes
Notes on term history	yes	yes (user-definable note type)	yes	no	yes (user-definable note type)	no	items from two user-definable lists ("reference" and "code") can be associated with each term and could be used for these purposes
Notes on term usage	no	yes (user-definable note type)	no	yes	yes (user-definable note type)	no	
Origin/authority for terms	yes	yes (user-definable note type)	no	yes	yes (user-definable note type)	no	
Term status (e.g. candidate)	no	yes	no	no	no	no	
Other predefined attributes	no	no	no	yes	no	no	no
User-defined attributes	no	4 note types, including SN	no	no	note types	no	no
Case preserved	yes	yes	yes	yes	yes	optional	inital capital is forced on all terms
Case significant	no	no (optional for search)	optional	no	no	no	optional
Symbols and spaces allowed in terms	not all symbols	yes	yes	yes	yes	yes, but spaces are replaced by underlines in printouts	yes
Name of package	**Beat**	**MultiTes**	**STRIDE**	**TAT**	**TCS**	**Term Manager**	**Hierarch**

Term relationships							
BT/NT, RT/RT, USE/UF	yes	yes	yes	yes	yes	yes	yes
USE X AND Y (semantic factoring)	multiple USE terms allowed but no special function	can be recorded and displayed using user-defined relationships such as USE_AND/USED_IN and USE_OR/USAGE_OF		no	yes, but searching for a "UF+" term gives "no match"	no	no
USE X OR Y (alternative terms)				no	no	no	no
Multiple BT links per term (polyhierarchy)	yes	yes	yes	yes	yes, but NTs shown under only one occurrence	yes	yes
No. of hierarchical levels	40	unlimited	100	9	unlimited	10	unlimited or user-defined limit
User-defined relationships	no	yes	yes	no	yes, for RT types only	existing relationships can be renamed but not added to	
Top term indicator	no	yes	no	yes	yes = "hierarchy name" (but this is not treated as a "term")	no	list of top terms can be displayed
Orphan term indicator	no	yes	no, though orphans may optionally be deleted when created by removing links	no	orphan terms can not be created	yes	list of orphan terms can be displayed
Multilingual thesauri	no (but handles and sorts accented characters according to the rules for Catalan and Spanish)	yes	no, though user-defined relationships may be used to link terms in different languages	yes	no	no	no
Name of package	**Beat**	**MultiTes**	**STRIDE**	**TAT**	**TCS**	**Term Manager**	**Hierarch**
Input and editing							
Import from file	yes	yes	yes	yes	can import a list of terms without relationships	yes	no
Input from keyboard	yes	yes	yes	yes	yes	yes	es
Editing of terms individually	yes	yes	yes	yes	yes	yes	yes
Consistency checks	yes	yes	yes	yes	yes	yes	yes
Loops detected	no	yes	no	yes	yes	no	yes
Terms multiply linked detected	yes	yes	yes	no (BT and RT allowed between same terms)	yes	no	no (BT and RT allowed between same terms)
Name of package	**Beat**	**MultiTes**	**STRIDE**	**TAT**	**TCS**	**Term Manager**	**Hierarch**
Output to screen							
Selection criteria for display of sub-set	no	yes	no subset display	yes (for alphabetical display)	no	wildcards	yes, extended list of wildcards and operators
Hierarchical display	no	yes	yes	yes (only NT relationships shown)	yes	no	yes (as screen display of a report)

Alphabetical display	yes	yes	no	yes (no relationships shown)	yes, but no relationships are shown and non-preferred terms are omitted	no	yes
Facet indicators	no	yes	no	no	yes	no	no
Sibling terms sorted in hierarchical display	yes (in single term display)	yes	yes	no	yes	no	yes
Classified (systematic) display	yes	yes	no	no	yes	no	no
KWIC or KWOC display	yes	yes	no	no	no	no	yes (as screen display of a report)
Scrolling and browsing	yes	yes	no	yes	yes	no	yes
Hypertext navigation (jumps between terms)	yes	yes. Also provides HTML output for use with a standard WWW browser	yes	no	no	yes	yes
Editing of terms while displayed as list	yes	no	yes, for single tree only	no	yes	no	no
Name of package	**Beat**	**MultiTes**	**STRIDE**	**TAT**	**TCS**	**Term Manager**	**Hierarch**
Output to printer or file							
Selection criteria for printing	yes	yes	yes (range of terms only)	no	not a range, but can limit hierarchical display to terms narrower than a chosen term	wildcards	yes, extended list of wildcards and operators
Hierarchical printout	yes	yes (two-way)	yes	yes (only NT relationships shown)	yes, full and brief formats	yes, but BT and RT are also shown at every level	yes
Alphabetical printout	yes	yes	yes	yes (no relationships shown)	yes	yes (same as hierarchical but showing one level only	yes
Facet indicators in printout	no	yes	no	no	yes	no	no
Sibling terms sorted in hierarchical printout	yes	yes	yes	no	yes	yes, if configured when thesaurus is first created	yes
Classified (systematic) printout	yes	yes	no	no	no	no	no
KWIC or KWOC printout	yes	yes	no	yes	yes	no	yes (both)
Export to file in a data exchange format	yes	yes	yes	yes	no	yes	yes

Name of package	**Beat**	**MultiTes**	**STRIDE**	**TAT**	**TCS**	**Term Manager**	**Hierarch**
Ergonomics							
Ease of installation	[4]	[4]	[3]	[3]	[4] Demo version gives no choice of destination directory	[3]	[4] if default locations are accepted
User interface	[3]	[4]	[3]	[3]	[3] Non-standard Windows controls	[3]	[3] Rather confusing because some controls apply to Hierarch and some to the underlying database, Paradox
Simultaneous creation of terms and relationships	yes	yes	yes	some	some	no	yes
Drag and drop relationships	no	no	no	no	no, though families can be moved by cut and paste	yes	no, but cut and paste is quite convenient
Printed documentation coverage	[0] (none included with shareware package)	[4]	[4]	[3]	[3]	[3]	[3] Mainly deals with technical issues of installation. Use of the software is documented in on-line help.
Printed documentation quality	[0]	[4]	[4]	[2]	[3]	[3]	[2] Assumes familiarity with thesaurus principles so no detailed user-level guidance.
On-line help	[3]	[3]	[3]	[0]	[3] Help file is the same as the printed manual. Doesn't use Windows Help: no hypertext or "find" functions. Doesn't work in dialog boxes.	[3]	[3] Two help systems, one for Hierarch and one for Paradox, not integrated. Context-sensitive help is not always available and only for Paradox when it is.
Name of package	**Beat**	**MultiTes**	**STRIDE**	**TAT**	**TCS**	**Term Manager**	**Hierarch**
General factors							
No. of terms	2 billion	100,000,000	unlimited	MS Access limit (?)	unlimited	64,000	unlimited
Related database available	no	Inmagic mentioned, but relationship not clear	has been integrated with STATUS	Possible links to Access; import/export to Tinlib	no	Cardbox for Windows	no

8.6. Metadata Tools Survey

The table below constitutes an analysis of available metadata tools on the US government FDGC last updated on the 25 July 2000.

Tool Name	Mac OS	Win 3.1	Win 95	Win NT	UNIX	Web Host	Fee?	Requirements	Import	Export	Software Info
ArcCatalog				X			Incl	ArcINFO 8.0.2	XML	XML	arccat.html
ArcView Metadata Collector	X	X	X	X	X		No	ArcView 3.X		mptext	csctool.html
ArcView Metadata Management System			X	X			Yes	ArcView 3.X		mptext	avmms.html
BIC Metadata Tool	W	W	W	W	W	UNIX	No	cgi-script	text	mptext, SGML	bic.html
CorpsMet95			X	X*			No	Standalone	mptext	mptext	corpsmet.html
DataLogr		X	X*				Yes	Standalone	text	text	
Data Dictionary (DataDict)				X*	X		No	ARC/INFO AML		mptext	datadict.html
Dataset Catalog Database Sys		X	X				No	Standalone	dbase	other	dcds.html
Document AML				X	X		No	ARC/INFO AML		mptext	document.html
Fgdcmeta AML				X	X		No	ARC/INFO AML		mptext	fgdcmeta.html
GeoData MDB			X	X*			No	Access 2.0		other	geodata.html
Geospatial Metadata Mgt Sys		X	X				No	Access 2.0		mptext	gmms.html
Metadata 2 (MD2)		X					No	Access 2.0		other	md2.html
Metadata Extension for ArcView	X	X	X	X	X		No	ArcView	text, SGML	html	meav.html
Metadata Entry System	W	W	W	W	W	UNIX	No	cgi-script	form entry	mptext, SGML	mes.html
MetaLite System for Windows			X	X			No	standalone	data entry	mptext, XML, SGML	mespc.html
Metadata Management System		X*	X				Yes	Standalone or DB	html	html	mms.html
MetaMaker 2.10		X	X				No	Standalone	other	mptext, dif	metamaker.html
Spatial Metadata Management System 2.0			X	X			Yes	Standalone	mptext, SGML	mptext, SGML, html, XML	smms.html
MetaStar (MDC, MDM, MDS)	J	J	J	J	J	UNIX NT	Yes	Standalone	SGML, custom	html, SGML, text, XML, custom	metastar.html
Xt Metadata Editor 1.9.1					X		No	Standalone	mptext, XML, SGML	mptext, XML, SGML	xtme.html

- * denotes that it may possibly work in that environment but behavior is unknown or undocumented.
- W denotes access through HTML Web browser.
-
- J denotes access through Java-based Web browser.