Libraries and Archives Collecting Newspaper Clippings

Unified for their Integration into Networks

LAURIN - LB-5629/A

Deliverable 3.4

Analysis and General Design of Indexing Systems

Version 1.2

1999-03-29

IZA - Institut für Germanistik Universität Innsbruck - Innsbrucker Zeitungsarchiv

Authors: Kurt Habitzel, Gregor Retti (IZA)

Contributors: Niko Hofinger (IZA) & Diego Calvanese (DIS)

Confidentiality level: public

Deliverable 3.4. Analysis of indexing systems

1. Executive Summary

Main questions of the analysis

These following questions framed the task 3.4 "Analysis of indexing systems":

- Which indexing and classification systems and methods are used by the LAURIN partners? Does their daily indexing work base on library standards?
- Which existing text-type indexes and subject indexes should be integrated into the LAURIN system?
- Are there existing thesauri systems or tools which can be re-used in the project? How do these systems handle multilingual thesauri?
- Are there feasible network solutions for a "distributed thesaurus"? How do other thesauri solve this problem?

Work done

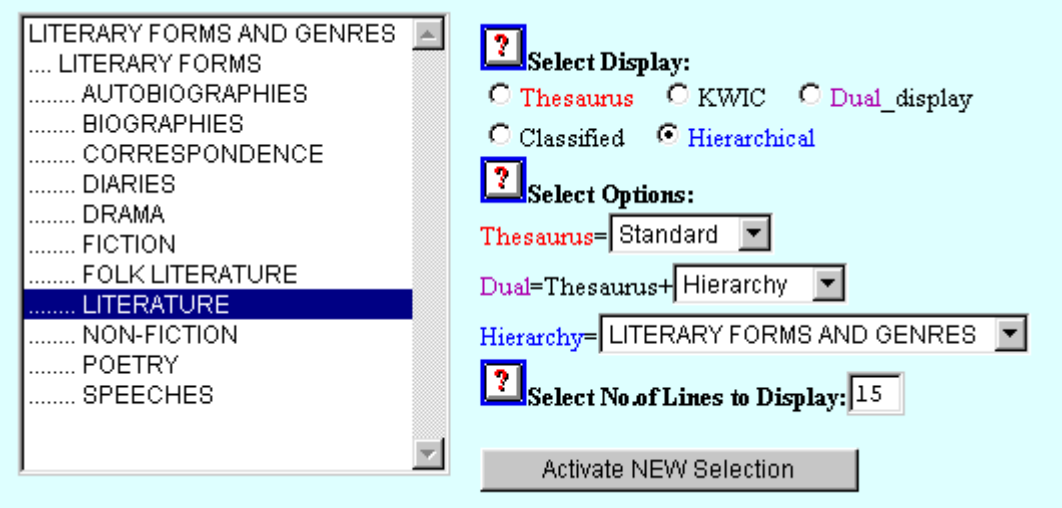
- The indexing systems of the participating libraries and archives were using a questionnaire. The answers given by the librarians can be found in Annex 3. Some additional inquiries by e-mail and contacts to other libraries (e.g. Literaturhaus Wien, Steiermärkische Landesbibliothek) completed these investigations.
- A sample of ten thesaurus software tools were tested. Software not available as trial version was evaluated with the help of information given on the Internet or by contacting the software vendors (brochures). A complete list of the tested or examined thesaurus tools is listed in Annex 2.
- Thesauri available on the WWW have been evaluated especially in regard of their user interface (display, capability to search and browse the terms) and the usability for the LAURIN-thesaurus (complete list see Annex 1). Additional other classification systems used by journalists or newspapers have been evaluated.
- A technical meeting with all other LAURIN developers was held in Rome in September in order to harmonise the technical approach.
- A paper on the functions of the LAURIN indexing system was published through the ftp-site of the project in order to give all partners the chance to give their feedback. Additional to this rather theoretical paper a mock up version of the future LAURIN indexing module was created in order to have a practical example to discuss on.

Main findings and conclusions

- The indexing work of most of the LAURIN archives does not base on library standards, as far as the aboutness of a clipping is concerned, while bibliographic data is recorded in rather similar ways. Most of them use "home-made" indexing systems which even don't exist in computer- or printed version. Nevertheless all LAURIN archives feel the need for standardisation and technical improvement. The LAURIN thesaurus will conform to ISO 2788-1986 (E) and ISO 5964-1985 (E).
- Six thesauri and classifications can be re-used for the LAURIN project (Licenses have been signed with the creators). Although these and other thesauri are valuable resources, clipping archives have to be flexible in assigning new terms: classification systems from the library world such as DDC, UDC etc. fit more the demands of organising collections of books. They cannot be applied to clipping archives and the world of journalism.
- Due to the fact that there is no appropriate software with a favourable price available for the maintenance of a multilingual, distributed thesaurus an own thesaurus-system based on a relational database has to be developed by the LAURIN project. This thesaurus-system will be closely integrated with the clipping database.

2. Analysis of existing thesauri

A number of thesauri, classification systems and subject headings are available in the Internet. They can be either downloaded (Adobe Acrobat-pdf or ASCII-files) or there is an easy access via a special WWW-interface.



Picture 1 HASSET - Humanities and Social Sciences Electronic Thesaurus

Most WWW-thesauri have two main navigation features for the user:

- search
  - term search
  - boolean search
- browse
  - alphabetical listings (sometimes starting with an alphabetical index)
  - alphabetical KWIC or KWOC (Keyword Out of Context) index listing
  - hierarchical browsing
  - hyperlinked hierarchical relations

Thesauri in two or more languages

Just a small number of the WWW-thesauri are multilingual (e.g. OECD Macrothesaurus, UDK-Online-Thesaurus, The Astronomy Thesaurus, CATIE Thesaurus, TGN). A wide range of languages are covered by EURODICAUTOM, the official site for the EU terminology. This database offers terms, translations, definitions and other term attributes in twelve European languages (see: http://www2.echo.lu/edic/).

Integration of existing thesauri and subject headings

Most of the thesauri available in the Internet cover highly specialised terminology. Therefore it does only make sense for a few of them to be integrated into the LAURIN thesaurus.

These are the following existing thesauri, classification systems and subject headings:

- We plan to use the IPTC (International Press Telecommunications Council, Windsor/United Kingdom) subject codes as top terms of the thesaurus system. The IPTC subject codes have been designed especially for the categorisation of news material. These codes comprise 345 terms (17 main subject names), which are hierarchically structured in 3 levels. The terms are available in English, French, Swedish, partly in Italian, and in Japanese, Croatian and Arabic. (See: http://www.xe.net/iptc/catguide.html). A draft implementation guideline is available, and we already have the permission of the IPTC to use these codes within our project.
- We plan to adopt parts of the macrothesaurus of the OECD (Paris) to broaden the IPTC subject codes. The macrothesaurus is a multilingual thesaurus (English, Spanish, French and partly German) and covers about 5000 terms especially in the fields of economic policy, industry, trade... It also comprises a list of international organisations. (See http://www-cui.darmstadt.gmd.de/~probst/thesa/).
- We already signed a license agreement with the Getty Information Institute, Los Angeles/USA, to use the TGN (Thesaurus of Geographic Names). This thesaurus contains about one million geographic names, has a multilingual structure and uses the vernacular names of cities, regions, countries, rivers, lakes etc. as preferred terms. It covers historical names of locations too. Diacritics are coded by a dollar character followed by two numbers (e.g. München = M$04unchen). The TGN was downloaded for LAURIN's purposes in summer 1998 (220 megabytes in size, relational file format).

  There is also a browsable version on the Internet (See: http://www.gii.getty.edu/tgn_browser/).
- We also have the permission to use the NUTS-Codes ("Nomenclature of Territorial Units for Statistics"), which have been developed by eurostat, Luxembourg and Brussels. These codes cover about 100.000 geographic names (=administrative districts: e.g. regions, cities, villages) from the countries of the European Union. We will use this hierarchical structured list to expand the TGN.

Hierarchy in the NUTS-codes is expressed by structure and notation. Structural hierarchy means that all geographic units are part of the broader geographic unit above them. Notational hierarchy is expressed by length of notation. Numbers at level 1, 2, 3 and 4 are subordinate to a class whose notation is one digit shorter. Therefore it will be possible to transfer the NUTS-code into the hierarchical structure of the LAURIN thesaurus and to combine it with the TGN.

Nuts	Code	Name	0	1	2	3	4	5
AT		Österreich	1	0	0	0	0	0
AT1		Ostösterreich	0	1	0	0	0	0
AT11		Burgenland	0	0	1	0	0	0
AT111		Mittelburgenland	0	0	0	1	1	0
AT11100001	10801	Deutschkreutz	0	0	0	0	0	1

Picture 2 NUTS

- We have signed a license agreement with the Getty Information Institute to use the ULAN (Union List of Artist Names), which includes 200.000 names representing approximately 100.000 individual artists. ULAN is delivered in a single ASCII text file with records in alphabetical order (about 30 megabytes storage). There is also a browsable version in the Internet (See: http://www.ahip.getty.edu/ulan_browser/).
- We also plan to adopt the writers database, which is maintained by the IZA and covers more than 30.000 proper names of authors and writers (See: http://iza.uibk.ac.at/db/).

So far no decission has been made concerning the field of economy. One option is the integration of the Standard Thesaurus Wirtschaft (See: http://www.hwwa.uni-hamburg.de/iz/thes.htm), which is available in German.
There are also some useful systems to classify branches in trade and products. The F.A.Z. (Frankfurter Allgemeine Zeitung) press archive uses the SIC code (U.S. Standard Industrial Classification), which was released by the US. Census Bureau. The North American Industry Classification System (NAICS), which was released in 1997, is replacing the SIC (See: http://www.census.gov/pub/epcd/www/naics.html).

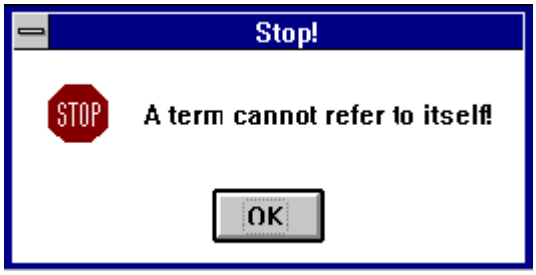
Analysis of existing thesaurus software

The objective of this task was to download a range of thesaurus-software (free of charge trials or demo versions) and to evaluate the functionality. Other important thesaurus management systems have been evaluated by the information we received from the software companies.

Almost all downloaded thesaurus tools (*adlib, BEAT, cardbox, cindex, dtsearch, hierarch, MultiThes, stride, TAT, TermTree*) are running on the MS-windows operating system. Two main LAURIN requirements, support of multilingual thesauri and adoption in a network, are not covered by most of the tested thesaurus tools.

The main features of the tested thesaurus tools are:

- support of the relationship types as specified in the ISO standard 2788 (USE/UF, BT/NT and RT)
- different types of listings and thesaurus reports
- automatic creation of inverse relationships
- Testing of the coherence of any modification in terms or relations (e.g. cross reference control, check for circular references)



Picture 3 MultiThes

Only a few thesaurus tools offered:

- an easy export format (e.g. HTML reports, platform independent export formats)
- an easy, user-friendly interface (e.g. using the relational structure of a thesaurus for navigation - "hyperlink")
- the possibility to create other relations than those specified in the ISO standard 2788
- administrative functions: keeping track of changes made in the thesaurus (who changed a term, when was it done...)

The tested software:

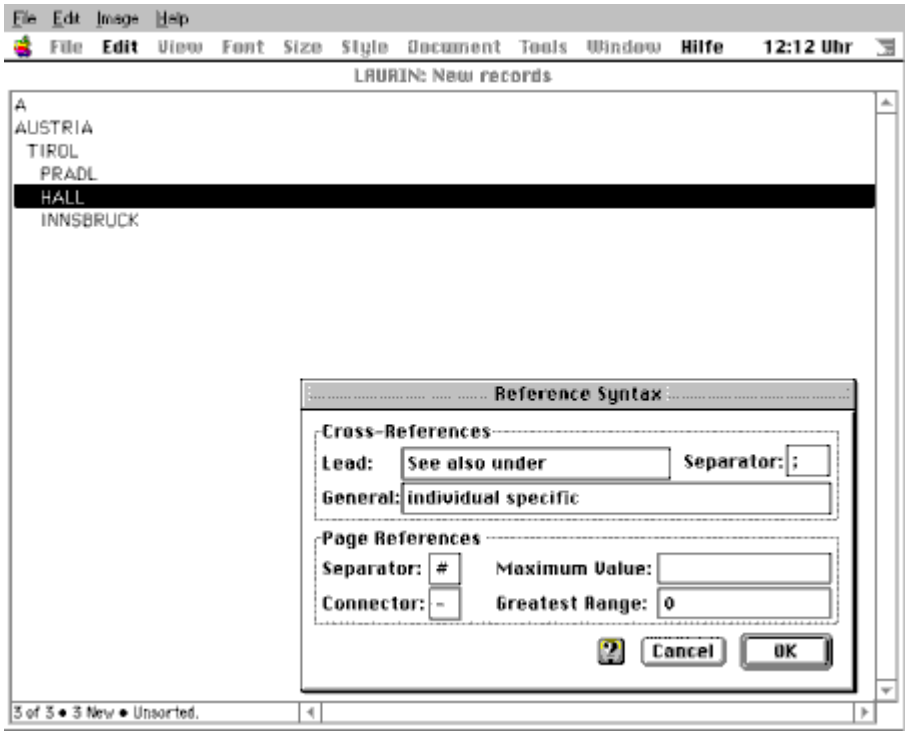
Not all thesauri tools are stand-alone software, some of them are modules of database packages (e.g. *STAR Thesaurus*, not tested), others have the functionality of a file management system (e.g. *dtsearch*) or are designed as library catalogue programme (*cardbox*, *adlib*) including some "thesaurus look-alike" indexing and retrieval functions. Therefore basic thesaurus software features are missing and the maintenance of a real ISO standard thesaurus is not possible.

TermTree is a simple tool for displaying hierarchical term lists and offers no additional thesaurus functions. *TAT*, a thesaurus-software using a MS-Access runtime module, did not work on our Windows 95 platform.

Therefore only five "real" thesaurus tools remained for a detailed testing:

BEAT is a MS-DOS based thesaurus software, supports English, Catalan and Spanish and other ISO Latin 1 characters and is easy to use with the keyboard. It offers many types of reports and display (e.g. hierarchical list, permuted keyword list, alphabetical list, rotated list). *BEAT* supports the relationship types as specified in the ISO standard (USE/UF, BT/NT and RT). It is possible to connect every term with scope notes, history notes and source notes. It is also possible to add a notation scheme. BEAT is free of charge.

CINDEX, a software also available for the Apple Macintosh, offers all necessary thesaurus features. It automatically corrects references following changes, verifies cross-references, tracks date and time when entries are added/edited. Additional features are spell-checking (only US and British English) and a wide range of import and export formats. Besides the conventional alphabetical listing it has a very simple hierarchical display.



Picture 4 CINDEX

The *Hierarch Thesaurus Manager* is a standalone Windows thesaurus management package. It includes circular references check, duplicate terms check, duplicate links check and forbidden term/allowed term check. A "relationship window" shows all information about a particular term in the thesaurus (BT, NT, RT, SN, categories, stop terms). This "trial software" is available in a demo movie only, so a detailed check could not be made.

MultiThes is a cheap and powerful thesaurus software. *MultiThes* was used to create the Thesaurus of Geographic Feature Type Terminology (see http://www.alexandria.ucsb.edu/~lhill/html/index.htm). It can administer monolingual and multilingual thesauri. Besides the standard features, *MultiThes* supports user defined relationships and comment fields. A HTML file generator creates all the files which are necessary to put a thesaurus on the Internet. A wide range of reports and displays can be created with this thesaurus tool.

Alphabetical Report

```
Alabama
  ABB:  AL
  BT:   USA
  NT:   Birmingham
        Mobile
        Montgomery
```

Hierarchical Report

```
Alabama
 ABB:  AL
 BT1:  USA
  BT2:  North America
   BT3:  America (continent)
    BT4:  Planet Earth
 NT1:  Birmingham
 NT1:  Mobile
 NT1:  Montgomery
```

Top Term Report

```
Planet Earth
. America (continent)
. . North America
. . . USA
 . . . . Alabama
 . . . . . Birmingham
 . . . . . Mobile
 . . . . . Montgomery
```

Rotated Index

```
AL
  AL
     ABF: Alabama
Alabama
  Alabama
```

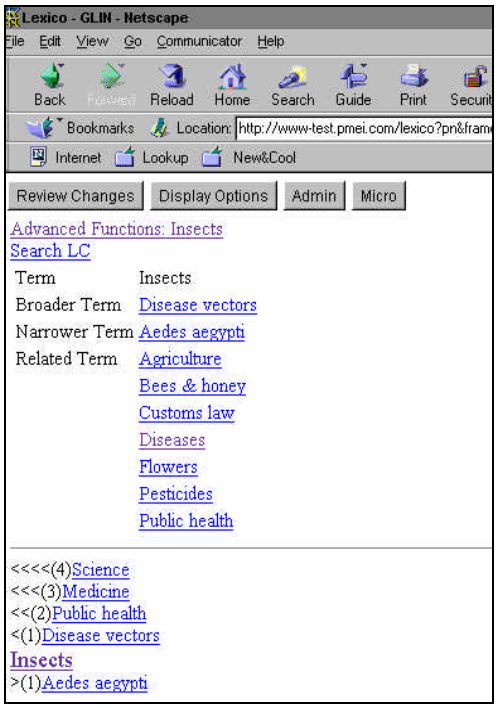
Picture 5 MultiThes

Stride is a multi-user system and supports distributed thesauri. *Stride* allows the use of all the standard relationships and the user can define any relationship that may be needed for use in special application areas. The programme has hypertext facilities which enable the user to shift to related contexts. Another feature is the possibility to define special views, where relationship symbols instead of relationship names are used. So a hierarchical view can be shown either in a more conventional text-based form or with tree lines (See below).

Picture 6 Stride

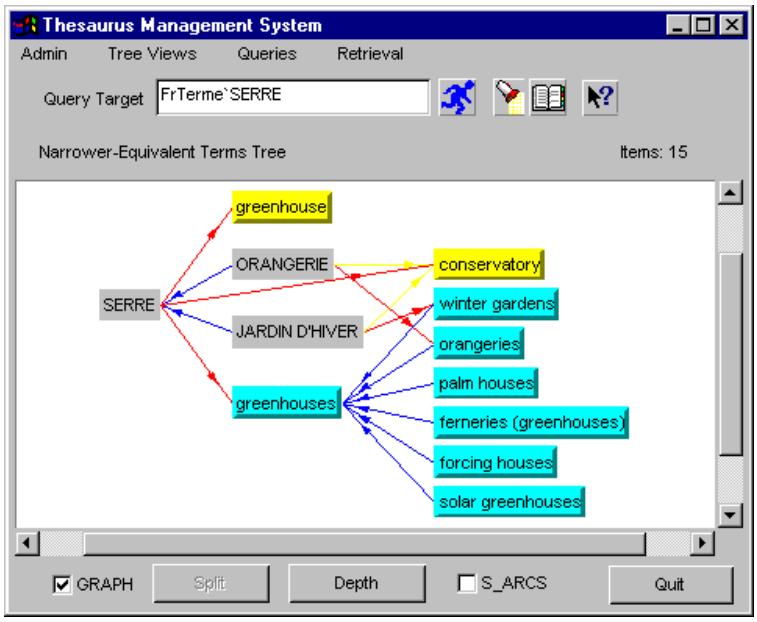
Other relevant thesaurus software:

LEXICO is a state-of-the-art thesaurus management system, expressly designed to handle the creation, maintenance and printing of automated vocabularies. The thesaurus can be displayed (printed) in alphabetical, hierarchical or KWOC (keyword out of context) format. An important feature for the LAURIN evaluation is that a LEXICO-maintained thesaurus can be accessed and manipulated over the Internet (Java-based interface that can be run with any browser). Lexico, which is available for a wide range of operating systems (Win95, WinNT Workstation, NT LAN Server, SUN SOLARIS, SGI IRIX, IBM AIX) is used in some important thesaurus projects (e.g. LC Thesaurus for Graphic Materials II, The Astronomy Thesaurus, Nuclear Regulatory Commission Thesaurus). The main disadvantage of LEXICO is the high price of 6000 $ for one license only.



Picture 7 Lexico

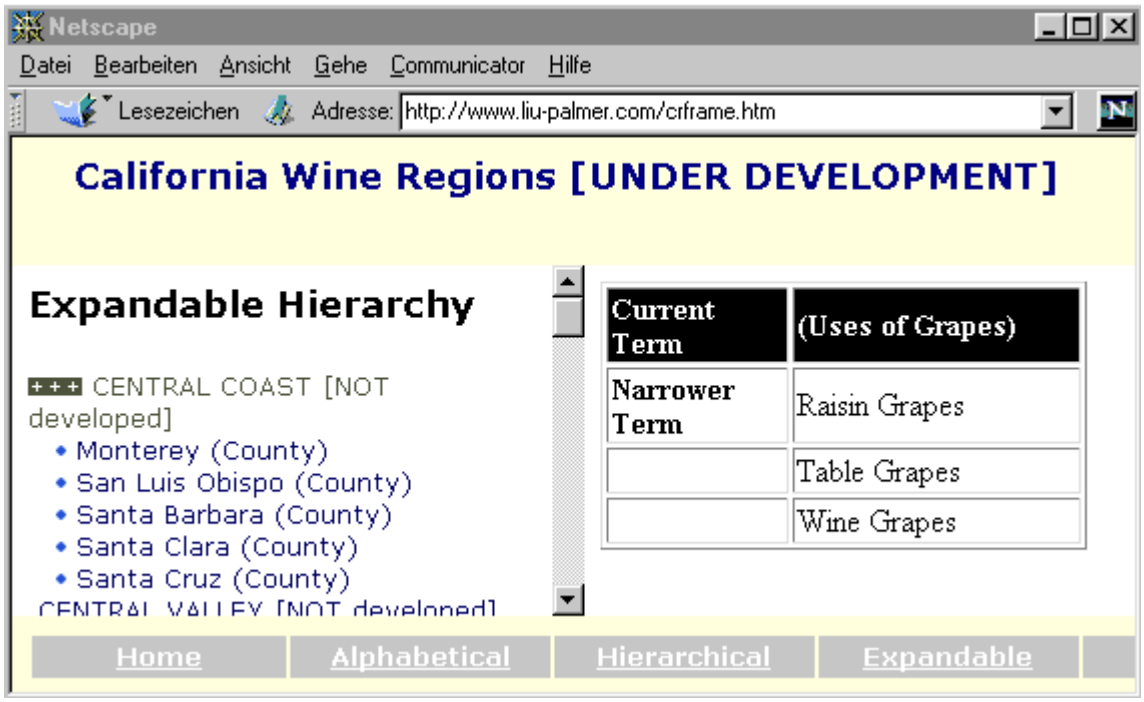
SIS-TMS is a thesaurus software developed to manage multilingual term relations. It is also a tool for a distributed access to heterogeneous electronic collections and to administrate a distributed thesaurus. It offers a graphical user-interface, which allows an easy navigation within the database. These features and the wide range of platform support (Win95/NT, Solaris, HP-UX, AIX) would make this software a feasible tool for LAURIN's needs. Despite the developer's announcements (ICS-FORTH, Crete/Greece) of a definite release this summer the software still remains in its beta version.



Picture 8 SIS-TMS

STAR/Thesaurus works with the information management system STAR only. The software offers all features necessary for thesaurus management including a support for notation. Together with STAR's Web interface a hyperlink navigation in the "virtual" thesaurus is possible. The main disadvantages are its price and the fact that it is an option of the STAR system database.

Thesaurus Construction System (TCS) and *Thesaurus Navigator 2000* (Liu Palmer).
The Thesaurus Construction System is a tool to build and maintain controlled vocabularies. One interesting feature is the possibility to move blocks within or across hierarchies.
The Thesaurus Navigator 2000 is a system that transfers TCS-thesauri into HTML-formats for access via WWW. A reasonable price does not compensate for the subservience to the Windows platform.



Picture 9 Thesaurus Navigator 2000

3. Analysis of the indexing and classification systems of the LAURIN partners

Subject headings, classification systems

The LAURIN project comprises 6 different clipping collections from 6 European countries. Some of these collections are basing on private initiatives and have therefore no "librarian" background. The Baldini collection was founded by Paolo Monelli, an Italian journalist, and the clippings are classified by his own home-made categories. The collections of IZA (founded by M. Klein) and ALV (founded by Th. Anz) started as private collections too. The archivation of clippings by CDP is of minor importance in the daily work, due to the fact that CDP is specialised in the production of present-day-press reviews. All four archives are not using library standards for their indexing work.

Only the Scandinavian LAURIN partners are using national library standards for their indexing work. The NBR is using the AACRII standard, the indexing work of UUL bases on the Swedish national library standard (SAB), which was adapted for the needs of the clipping collection.

Most of the libraries use singular and plural forms for the indexing terms. They also use abbreviations and multi-term keywords. Normally there is no distinction of homonyms, only the UUL uses qualifiers. No library uses scope notes or definitions for a detailed description of the subject headings.

All participating archives use their national languages for indexing. Apart from CDP, no library/archive is using vernacular names for geographic terms. A main problem is the handling of transliterations from non-roman alphabets: Most of the libraries have no common rules, sometimes (like CDP) the handling depends on the origin of the source.

No library uses a known classification like DDC for their collection. UUL uses a "home-made" classification, the Baldini collection is organised by the historical "classification" of Paolo Monelli, and CDP/UOC are using an index according to the handled themes.

Text types

The IZA, UUL, NBR and CDP/UOC are classifying their articles according to the text types used. Only IZA and UUL provided a list of these text types in English. Due to the fact that the interest of IZA is in the fields of literature and UUL has a much broader collection, both lists are completely different. The following table shows the text types provided by IZA and UUL. To get a wider range of text types we added in the "text types" (="attributes of an news object") by the IPTC (International Press Telecommunications Council). The IPTC-attributes describe the nature or characteristics of a news object, not specifically its content.

UUL	IZA	HWWA	IPTC	
			Object Attribute Name	**Object Attribute Description**
editorial book review news article feature article commentary speech interview	book review review on theatre, broadcast, film performance article on awards biographical article announcement report about events article on "memorial days" primary text, such as essay, poem,.. other	ranking forecast history Documentation Opinion Review Fair Conference Interview Biography Information on Companies Branches/trades Products Markets Countries	Current	Indicates the information is about events taking place at the time of the report.
			Analysis	Data and conclusions drawn by a journalist who has researched the story in depth.
			Archive material	Material distributed previously that has been selected from the originator's archives.
			Background	Provides some scene setting and explanation for the event being reported.
			Feature	The information is about a particular event or individual that may not be significant to the current breaking news.
			Forecast	Used when the Object contains opinion as to the outcome of a future event.
			History	Material based on previous rather than current events.
			Obituary	A narrative about an individual's life and achievements for publication after his or her death.
			Opinion	An editorial comment that reflects the views of the author.
			Polls & Surveys	Numeric or other information produced as a result of a questionnaire to a sample of the population.
			Profile	A narrative about the life and achievements of a living individual.
			Results Listings & Tables	Numerical data presented in tabular form for easier understanding.
			Side bar & Supporting information	A related story that provides additional insight into the news event being reported.
			Summary	A number of stories that have been reduced in length and compiled into a single news item.
			Transcript & Verbatim	The written version of an interview or uttered Statement without alteration or comment.

4. General design

Indexing

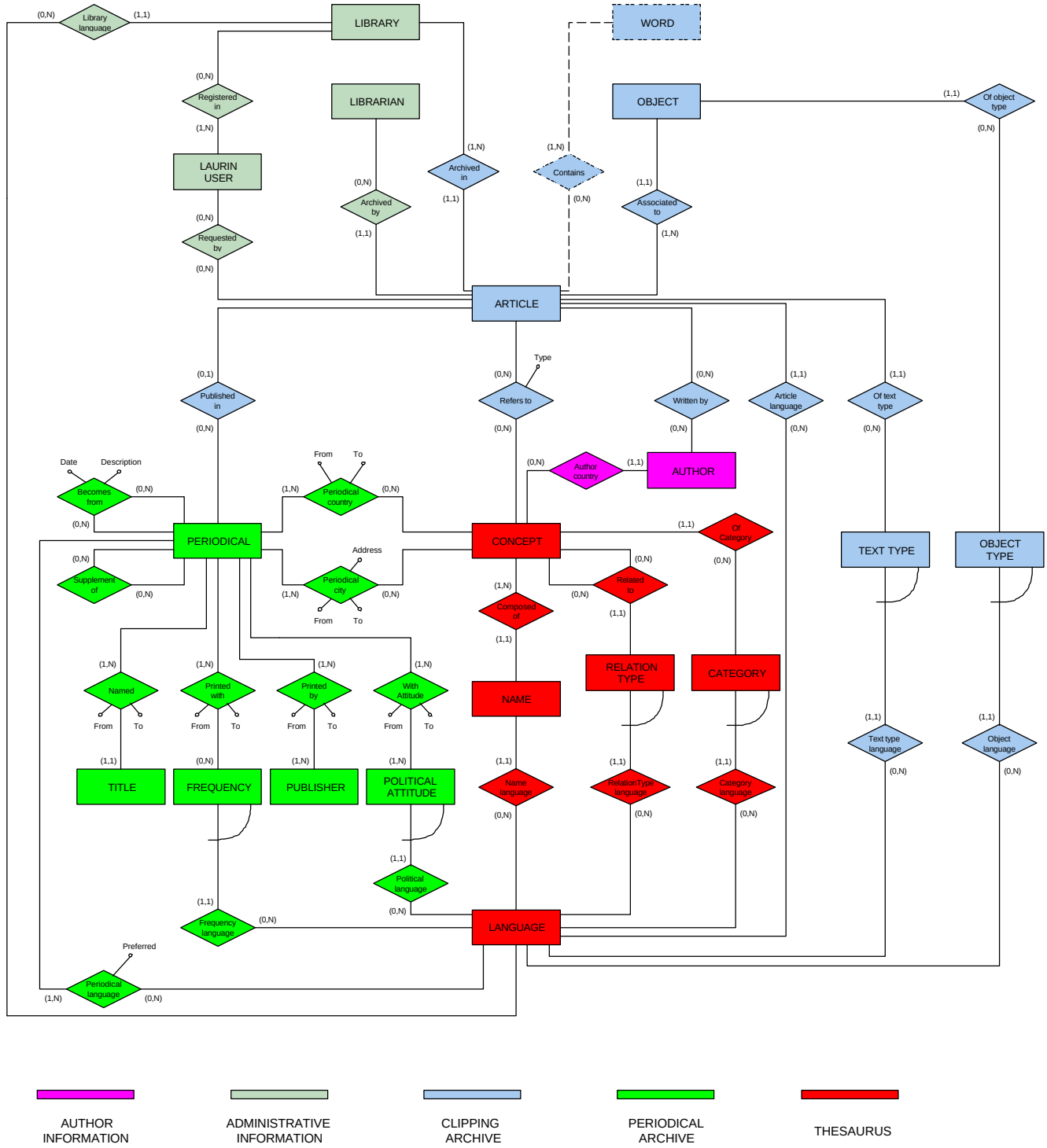
The LAURIN-system will contain the following six indexes.
Some of them will be generated automatically, others have to be filled and maintained by the librarians:

- **Prime index**: Basic information which otherwise will be lost during the clipping process (name of newspaper, page, rubric, date,...).
- **Bibliographic index**: Basic bibliographic information (author, title, subtitle, text type of an article).
- **Keyword index**: Computer based association of known terms from the thesaurus with clipping/article.
- **Content index**: Association of clipping/article with normalised term(s) from the thesaurus using the keyword indexing and additional associations derived from human content analysis of the clipping/article.
- **Free index**: Association of clipping/article with subject headings that are not part of the thesaurus. The terms in the free index are candidates for the thesaurus.
- **Full-text index**: Computer based retrieving of all normalised terms in the clipping (generated by a full-text information retrieval engine).

Technical solution for "content indexing" (thesaurus)

The thesaurus will be organised as a set of relational tables and stored in a relational database. The LAURIN-thesaurus is organised by concept. The purpose is to link alternative names in different languages of the same concept together and to identify useful relationships between these different concepts. Every *concept* will have a *unique key* and every *concept* will be represented by several *names* (including name string, normalised name string, language flag, preferred flag).
E.g. the unique key 1234 (representing <house>) will be represented by several names: "casa" (Italian), "Haus" (German), "house" (English),... and by the information on which name is the preferred one.
The thesaurus will also contain the information about the relationships between concepts (e.g. broader term, related term,...) and some administrative information (who changed/added when and what in the thesaurus).

The basic structure of the tables that constitute the thesaurus with their most relevant attributes are shown in the ER-figure (cf D3.2, CM-Sistemi):



Picture 10 ER

The precise set-up of the tables that constitute the thesaurus will be determined on the basis of a more detailed analysis of the types and occurrences of relationships between concepts. The thesaurus will be mainly used for *Content Indexing*, which offers the possibility of a controlled and normalised indexing and for a multilingual retrieval. The maintenance of the thesaurus will be done by the librarians, one of them (during the project a person at the IZA) will have the position of a supervisor.

Each article will be associated to a set of concepts which are contained in the thesaurus (This association is in fact the association of the unique keys of the *concepts* and the unique key of the *article*).

If it is necessary to apply a new concept (which is not in the thesaurus yet) during content indexing, this new concept can be added to the "free vocabulary" (free indexing). Such a concept has to be regarded as a candidate for insertion to the thesaurus. Upon validation of the concept and insertion to the thesaurus the "free index" entry of the article that refers to the concept becomes a "*content* index" entry.
Maintenance of the thesaurus will take place in a clear and well-structured way. Therefore workflow, case-studies, technical as well as organisational solutions have to be worked out and provided.
A rough scheme of this workflow may look like this:

- a new concept is identified during content indexing and stored as free vocabulary
- the local thesaurus manager (i.e. the responsible person for thesaurus maintenance at a local node) checks the free vocabulary entry. He may find that the concept already exists, that it is a new name for an existing concept or that it is a new concept or he may reject it. To decide upon this question he may use the article references by the free vocabulary entry. If it is a new concept he must provide a set of additional information (e.g. source, language, suggested relations, English translation and/or English scope note).
- The new concept is then sent to the central node and will be validated there by the thesaurus supervisor.

The central idea of a workflow like this is to clearly distinguish between content indexing and thesaurus maintenance. Content indexing is the only stage where new concepts can be captured. However it is still part of the acquisition workflow and should therefore not be overloaded with other tasks.

Text types

Adding information about the text type of an article is regarded to be an important feature. The answers gathered through the questionnaire on this subject and from other sources (e.g. HWWA, IPTC, RSWK) did not produce a coherent list of text types. Different points of view as well as different aims of the collections determine the "text type" used by the archives. Further investigations have to be made to compile a list suitable for the classification of the text type of the articles. At the time being it is foreseen to offer a text type classification on the level of the **bibliographic index**. A classification seems to fit better the needs for this feature than a conception which inherits the structure of a thesaurus. A multilevel classification should be flexible enough to cover the different points of view and aims of the archives (e.g. http://germanistik.uibk.ac.at/menorah/desc.html#ts). "Text type" in a broader sense may cover different types of images as parts of the articles too. Anyway this classification will have a rather limited number of elements and therefore it will be easy to provide it in all Laurin-languages.
A first step to build up such a list has shown, that the usage of semantic markers is very helpful to describe and distinguish different text type according to their dimensions (e.g. +/- current; +/- opinion).

Functionality of the thesaurus management system

The thesaurus management system will contain the following features:

- a user interface for thesaurus administration
- access rights management for different types of users (librarian, local thesaurus manager, central thesaurus supervisor)
- log of all changes in the thesaurus (who what when)
- keeping track of all concepts/names, which have to be revised in the thesaurus maintenance workflow
- consistency check (e.g. check for missing references or circular references or orphan concepts; discouraged concepts cannot be made preferred concepts)

References

- Atchinson, Jean; Gilchrist, Alan: Thesaurus construction. A practical manual. London: Aslib 2nd ed. 1987.
- Austin, Derek: Vocabulary control and information technology. In: Aslib Proceedings, 38 (1), January 1986, p. 1-15
- Chaumier, Jacques: Le traitement linguistique de l'information. Paris: Entreprise Modern d'Édition. 3e édition. 1988.
- Erstellung und Weiterentwicklung von Thesauri. Einsprachige Thesauri DIN 1463, Teil 1 1987.
- Erstellung und Weiterentwicklung von Thesauri. Mehrsprachige Thesauri DIN 1463, Teil 2 (Entwurf) 1988.
- Harping, Patricia et. al.: User's Guide to TGN: Relational Files Format, Version 1.0. Los Angeles: Getty Information Institute 1998.
- Das F.A.Z.-Datenbank-Handbuch. Suchhilfen für die Datenbanken der Franfurter Allgemeinen Zeitung online und auf CD-ROM. Frankfurt a. M.: F.A.Z. 1. Auflage 1998.
- Ganzmann, Jochen: Check-list for thesaurus software. http://www.willpower.demon.co.uk/CriteriaFrames.htm (Originally published in: International Classification, 1990, vol. 17, no. 3/4, p. 155-157)
- International Terminology Working Group: Guidelines for Forming Language Equivalents: A Model Based on the Art & Architecture Thesaurus. http://www.gii.getty.edu/guidelines/index.html
- International Standards Organization (ISO). Documentation - Guidelines for the Establishment and Development of Monolingual Thesauri. ISO 2788-1986 (E). Geneva: ISO 2nd ed. 1986.
- International Standards Organization (ISO). Documentation - Guidelines for the Establishment and Development of Multilingual Thesauri. ISO 5964-1985 (E). Geneva: ISO 1st ed. 1985
- International Standards Organization (ISO). Documentation - Bibliographic references - Content, form and structure. ISO 690: 1987 (E), Geneva: ISO 2nd edition 1987
- Introduction to TGM I. http://lcweb.loc.gov/rr/print/tgm1/ (2nd ed 1995)
- Introduction to TGM II. http://lcweb.loc.gov/rr/print/tgm2/ (2nd ed 1994)
- IPTC - NAA. Information Interchange Model Guidline 3. Paris: Comité International des Télécommunications de Presse
- Janus, Bridget: The Clipping Collection. In: News Media Libraries. A Management Handbook. Ed. by Barbara P. Semonche. Westport 1993.
- Komorous, Hana; Harriaman, Robert B.: International guidelines for the cataloguing of newspapers. London: UBCIM Programme (=IFLA Universal Bibliographic Control and International MARC Programme) 1989
- Di Lauro, Anne: Guide to the Maintenance of the Marcrothesaurus for Information Processing in the Field of Economic and Social Development. Paris: OECD 1993.

- Lutes, Barbara: Web Thesaurus Compendium. http://www.darmstadt.gmd.de/~lutes/thesauri.html
- Macrothesaurus. Multilingual Thesaurus Management and Term Retrieval System. User Documentation. Version 3.1. Paris: OECD 1997.
- Maniez, Jacques: Relationships in Thesauri: Some Critical Remarks. In: International Classification 15 (1988) No. 3. pp. 133 -138.
- Noelle-Naumann, Elisabeth et. al.: Publizistik. Massenkommunikation. Das Fischer Lexikon. Frankfurt a. M.: Fischer Taschenbuch Verlag 1989
- Regeln für den Schlagwortkatalog RSWK. Bearbeitet von der Kommission des Deutschen Bibliotheksinstituts für Sacherschließung. Berlin: DBI 2. erweiterte Auflage 1991.
- Roberts, Norman: The Pre-History of the Information Retrieval Thesaurus. In: Journal of Documentation, Vol. 40, No. 4, Dec. 1984, pp. 271-285
- User's Guide to ULAN:REC, Version 1.0. Los Angeles: Getty Information Institute.
- Weinberg, Bella Hass: Complexity In Indexing Systems -- Abandonment And Failure: Implications For Organizing The Internet. ASIS Annual Conference Proceedings October 19-24 1996. http://www.asis.org/annual-96/ElectronicProceedings/weinberg.html
- Will, Leonard: Thesaurus principles and practice. http://www.willpower.demon.co.uk/thesprin.htm (Revised version 13th February 1998.).
- Yuan, Quinming; Chang, Ifay: IT thesaurus construction - the methodology and observations. http://pride-i2.poly.edu/~qmyz/papers/iasted/ias_pap.html
- Zimmermann, Harald H.: Anmerkungen zur Neugestaltung der DIN 1463 (Thesauri). In: Fortschritte in der Wissensorganisation 2 (1992). S. 313-318.

Annex 1: Thesauri and classification in the WWW

- The Art & Architecture Thesaurus (Getty Information Institute)
  http://www.gii.getty.edu/aat_browser/
  Access/Display:
  Hierarchical listing (HTML), hierarchical navigation, search interface
  120,000 terms for object, textural materials, images, architecture and material culture description
- The Astronomy Thesaurus (International Astronomical Union)
  http://msowww.anu.edu.au/library/thesaurus/
  Multilingual (English, French, German, Italian, Spanish)
  Access/Display:
  Alphabetical and hierarchical listing (HTML)
  4000 terms in the field of astronomy
  Uses LEXICO thesaurus software
- CALL Dictionary and Thesaurus (US Government)
  http://call.army.mil/call/thesaur/index.htm
  Military terminology and personal names
  Access/Display:
  Alphabetical browsing, search interface
- CATIE-Thesaurus (Canadian "Community AIDS Treatment Information Exchange")
  Access/Display:
  Key words out of context (KWOC)
  In English and French
- CAS General Subject Vocabulary Helper, 14th Collective Index Period (Chemical Abstracts Service/CAS - American Chemical Society)
  http://www.cas.org/vocabulary/index.html
  terms in the field of chemistry
  Access/Display:
  Alphabetical listing
  Hierarchical display (user chooses the hierarchical depth to display)
- Common Procurement Vocabulary (CPV) (European Commission)
  http://ted.eur-op.eu.int/en/cpv_en.html
  5600 terms in 11 European languages
  Classification list (pdf, rtf and doc), hierarchically structured
- Dewey Decimal Classification, 21st version (OCLC)
  http://www.oclc.org/oclc/fp/about/ddc21sm1.htm
  30 languages, about 1000 "terms"
  Structural and notional hierarchy
  Access/Display:
  Hierarchical browsing

  or
  WWlib Browse Interface
  http://www.scit.wlv.ac.uk/wwlib/browse.html

- Draft Thesaurus of Information Object Type Terminology (Alexandria Digital Library)
  http://www.alexandria.ucsb.edu/~lhill/objtype_html/index.htm
  - 100 terms
    - Alphabetical listing
    - Hierarchical listing of top terms
    - Tagged text listing
- Global Legal Information Network (GLIN) Thesaurus - Library of Congress
  http://lcweb2.loc.gov/glin/indxhlp.html
  - 54,000 terms in the field of legislation
  - Access/Display:
    - hyperlinked terms, hierarchical display
    - browsing and searching
- HASSET (Humanities And Social Science Electronic Thesaurus) Version 2.0 - University of Essex
  http://155.245.254.46/services/zhasset.html
  - Access/Display:
    - Thesaurus, KWIC, Classified
- The ICONCLASS Browser
  http://iconclass.let.ruu.nl/
  - 13,000 unique keywords on iconographic
  - Access/Display:
    - classification system (HTML and experimental Java version)
- MeSH98Subject Headings
  http://omni.ac.uk/umls/
  - Fields of medicine, basing on the UMLS Metathesaurus
- NASA Thesaurus (1998 Edition)
  http://www.sti.nasa.gov/thesfrm1.htm
  - Access/Display:
    - Hierarchical listing with definition (pdf)
    - Hierarchical listing (ASCII File)
    - Rotated term display (pdf)
  - 17700 terms
- Nuclear Regulatory Commission Thesaurus
  http://www.pmei.com/nrc/
  - Access/Display:
    - Alphabetical listing, hierarchical display
  - Uses LEXICO/2 thesaurus software
- OECD Macrothesaurus (OECD)
  http://info.uibk.ac.at/info/oecd-macroth/en/index.html or
  http://www-cui.darmstadt.gmd.de/~probst/thesa/
  - Access/Display:
    - Alphabetical listing, search interface
  - About 5000 terms (specialisation on economic policy, industry, trade,...)
  - Multilingual (English, French, Spanish and partly in German)
- Getty Thesaurus of Geographic Names (Getty information institute) -
  http://www.gii.getty.edu/tgn_browser/
  - Access/Display:
    - Hierarchical listing (HTML), hierarchy navigation, search interface
  - One million geographic names, multilingual, using the vernacular names

- LC Thesaurus for Graphic Materials I. Subject matter of graphic materials
- LC Thesaurus for Graphic Materials II: Genre and Physical Characteristic Terms (TGM II), Library of Congress
  http://lcweb.loc.gov/rr/print/tgm1
  http://lcweb.loc.gov/rr/print/tgm2
  5,500 terms (TGM I) and 600 terms (TGM II)
  Uses LEXICO thesaurus implementation software, searching and browsing
- Universal Decimal Classification - Antwerpen Library
  http://www.ua.ac.be:80/MAN/UDC/udce.html
  Multilingual (English, Dutch, and French) classification for libraries
- UDK-Online-Thesaurus (Umweltbundesamt Vienna)
  http://udk.bmu.gv.at/thes/thes_acro.html
  Multilingual (English, German)
  Access/Display:
  Hierarchical listing (HTML), hierarchy navigation
  8500 terms about environmental information
- UMLS Metathesaurus (National Library of Medicine)
  http://www.nlm.nih.gov/pubs/factsheets/umlsmeta.html
  The Metathesaurus contains 476,322 biomedical concepts with 1,051,903 different concept names from more than 40 source vocabularies
  ASCII relational and Astract Syntax Notation (ASN.1) formats
  Documentation (http://www.nlm.nih.gov/research/umls/UMLSDOC.HTML)
- The Union List of Artist Names Browser (Getty Information Institute) -
  http://www.ahip.getty.edu/ulan_browser/
  Access/Display:
  Alphabetical browsing (HTML) including cross-references from non-descriptors (variant names)
  Search interface
  200,000 names representing approximately 100,000 individual artists
- WordNet - Lexical Database for English
  http://www.cogsci.princeton.edu/~wn/
  Lexical database for English, relations link the synonym sets
- Plumb Design Visual Thesaurus - WordNet
  http://www.plumbdesign.com/thesaurus/
  Access/Display:
  Java based visual interface (spatial map of relations)

Annex 2: Thesaurus software

Title	URL	demonstration available	OS	documentation	user or thesauri
Cindex	http://www.indexres.com/cindex.html	Windows	Windows 32-bit, Mac	winguide.pdf	
dtSearch	http://www.dtsearch.com/	Windows	Windows 32-bit	dtmanual5.pdf	
Stride	http://www.questans.co.uk/	Windows			
Speed	http://www.questans.co.uk/	Windows	Windows, DOS, UNIX		Wordnet.z (Encrypted)
STAR	http://www.cuadra.com/products/thesaurus.html	not available			
ICS-TMS	http://www.ics.forth.gr/proj/isst/Systems/TMS/index.html	not available Beta-stadium	Win95, WinNT, Solaris, HP-UX, AIX		AAT, u.a.
Multites	http://www.concentric.net/~Multites/	Windows Win95	Windows NT, or Win95, Windows 3.1	Lesson1.doc-Lesson7.doc	e.g. US-Army
LiuPalmer	http://www.liu-palmer.com/products.htm	partly online	UNIX; Windows NT		
Infologics	http://www.infologics.com/	no	Microsoft Windows NT		
Databasix: Adlib	http://www.dis.nl/ http://www.disbv.com/index.shtml	yes	UNIX, MS-DOS, Windows 3.x, 95 and NT		
BASIS	http://www.idi.oclc.org/html/basisv8.htm	no	UNIX, Windows NT		
BEAT THESAURUS SOFTWARE	http://www.zeta.org.au/~aussi/software/thesauri.htm	yes	DOS		
The Hierarch Thesaurus Manager	http://www.ozemail.com.au/~sisnsw/hierarch.htm	Lotus Screencam demonstration	Windows		
MTM für CDS/ISIS	http://www.oecd.org/dev/lib/macroa.htm http://www.unesco.org/webworld/isis/isis.htm	no	DOS, UNIX, Windows		
BiblioTech	http://www.bibliotechpro.com/thesaurus.html	no	WinNT, UNIX, Sun		
BRS/SEARCH	http://www.dataware.com/	no	Win95, WinNT,		

			UNIX, VMS		
*CAIRS	http://www.cairs.co.uk/	no	?		
Lexico	http://www.pmei.com/lexico/lexico.html	no	Win95, WinNT, Sun, Linux, Unix		
Lexico/2	http://www.pmei.com/lexico/lexico.html	no	OS/2		
Thesaurus Administration Tool (TAT)	http://www.psp.cz/kps/knih/eurovoc/papers/tat.htm	yes	Windows31	Tat-eng.doc	
TERM MANAGER	http://www.cardbox.co.uk/	yes	Windows		
TermTree	ftp://adam.ac.uk/pub/ADAM/tony/TermTree/ (not available anymore)	yes	Windows 31, Win95		
ThesMain	http://udk.ubavie.gv.at/thes/thes_acro.html	no THESmain is free of charge for all European Authorities	Windows31, Windows 95, Windows NT 3.51, Windows NT 4.0, OS/2		

Annex 3: Answers in the LAURIN questionnaire

Question	IZA	Baldini	ALV	UUL	NBR	UOC	CDP
Please mark the index characteristics you use; base-index	short source name, issue date	-	short source name, issue date	short source name, issue date	long source name, issue date, side number	long source name, issue date, side number, issue number	long source name, issue date, side number, issue number
Advanced bibliographic registration	-	-	-	title, author, heading	title, abstract, author	title, author	title, author
Content development (low level)	name of the persons, the lands, the institutions / companies, events concerned	-	name of the persons, events concerned, subjects	name of the persons, the lands, the institutions / companies, events concerned	name of the persons, the lands, the institutions / companies, events concerned, date of event, genre, registrated work, author	name of the persons, the lands, the institutions / companies, events concerned, thematic	name of the persons, the lands, the institutions / companies, events concerned, thematic
Full-text	no	-	?	no	yes?	yes?	yes?
Abstracts	no	-	no		yes	no	no
Do you use geographical names	yes, some hundred	-	no	yes	planned	yes	yes
Do you use proper names	yes, approx. 30000	-	yes,...	yes	approx. 8000	yes	yes
Do you use singular or plural form?	plural & singular	-	singular	singular + plural	not decided yet	singular + plural (sometimes)	singular + plural (sometimes)
Do you use articles?	no	-	no	yes, when appropriate	-	no	no
How do you normally use combinations of adjective + noun?	adjective + noun	-	only nouns	adjective + noun	not decided yet	adjective + noun	adjective + noun
How do you scope with multi-term keywords	multi-term keywords	-	preferred: Komposita	avoid to split keywords	not decided yet	education + adults	education + adults
Do you rather use ab-breviations or long versions?	no abbreviations	-	abbreviation	abbreviation	both, most used form	abbreviation	abbreviation

Question	IZA	Baldini	ALV	UUL	NBR	UOC	CDP
How do you combine multiple keywords? Do you use a specific syntactical order?	no	-	-	normal praxis of usage	not decided yet	yes, thematic, section, subsection	yes, thematic, section, subsection
How do you distinguish homonyms?	no distinction	-	we hardly have those in German	qualifiers	not decided yet	implicit	implicit
Do you use cross-references or any other means to relate subject headings / keywords?	yes	-	yes	yes	would like to do.	no	no
Do you use definitions / scope notes for your subject headings / keywords?	no	-	no	no	not likely	no	no
Do you have and use guidelines for your subject headings?	no	-	no	-	not yet	yes, thematic inclines	yes, thematic inclines
Which language(s) do you use for indexing?	German	-	German	Swedish	Norwegian	Catalan	Catalan
Are there any digital resources for your language(s)?	no	-	yes, but not at ALV	no	not sure	yes	yes
Which language do you use for geographical names and for proper names?	German	-	German	Swedish	Norwegian	Catalan	Catalan
Do you use the vernacular names?	no	-	no	no	-	yes	yes
How do you handle trans-literations from non-roman alphabets?	no rules, common use	-	no rules, common use	according to bibliographic rules	we only use one form of the name, but with references from all others	depends on original source	depends on original source

Question	IZA	Baldini	ALV	UUL	NBR	UOC	CDP
Do you use a known classification for your collection?	no	-	no	no	no	no	no
Do you use any other classification?	no	-	no	yes, home-made	no	yes, index according to handled themes	yes, index according to handled themes
Do you use a printed /computerised version of your indexing system?	yes, print-version	-	computer version	no	no	yes, partially	yes, partially
If you are using a computer to maintain your indexing systems, please describe in detail	print-version	-	Word for windows	relational database (CDS/ISIS) and MS Access / OpenImage, standalone and net	-	-	-
Do you classify your articles according to the "text types" used	yes	-	no	yes	yes	according to thematic	according to thematic
Do you have a printed (computerised) version of this classification / typology?	yes	-	no	no	yes	no	no
If so, please provide it in your language(s) and in English.	book rev., portraits, interviews, articles on awards, reviews on theatre, broadcast, film performance announce-ments reports about events, articles on memory days, essays, poems, others	-	-	editorial, news article, feature article, commentary, speech, interview, book review	-	-	-

Question	IZA	Baldini	ALV	UUL	NBR	UOC	CDP
Which library-standards do you use for your indexing-work in your newspaper / clipping-department?	no standards			Swedish national library standard (SAB), adapted to newspaper...	AACR II	other (proprietary)	other (proprietary)
Do you use other sources for your indexing systems, which might be useful as references?	no	-	-	yes, when needed	building up an own special dictionary	no	no
How and when is the indexing of the articles done?	manually, after pasting the articles and before copying them for multiple filing	-	-	long after the clippings are filed	after selecting articles	every 15 days and monthly	every 15 days and monthly
Please describe in detail the process of indexing in your archive.	one person, B.A. for German literature. Indexing after clipping 2-3 subject headings per article keywords including proper names up to 10	-	-	one person is indexing only 2 newspapers	index office is responsible. Librarian assistant, Indexing is done while selecting	analyst. University degree theme - sub-theme-section	analyst. University degree theme - sub-theme-section
Can you provide a rough estimation about the amount of your daily work for indexing compared to...	10%	-	-	2 hours	irrelevant	1h/3 persons	1h/3 persons
How is maintaining, updating of your indexing system organised?	The person who is indexing maintains the system	-	-	new words are added continually	IT department	paper, no organisation	paper, no organisation

Question	IZA	Baldini	ALV	UUL	NBR	UOC	CDP
What is the percentage of clippings you are not able to serve in indexing and cataloguing	-	-	-		irrelevant	10%	10%
How many are the old and new sources you are not indexing	none	-	-	many	-	none	none/some
Which is your ideal indexing schema	transfer old indexing system in new medium, handling should be browser orientated, thesaurus, Internet resources included, procedure should be transparent and simple, including of expert group	-	-	heading, source, date, article type, article size, author, named person, company, geographical name, link to other relevant articles, keywords	-	theme sub-theme section date author communi-cation, media text contained keywords	theme sub-theme section date author communi-cation, media text contained keywords
Do you need to increase the % of usage of your computer-based indexing system?	yes	-	-	yes	-	yes	yes
Do you need to increase your PC usage, LAN services usage, WAN services usage, during such activity?	yes	-	-	yes	-	yes	yes
Do you need an enhancement of the system interface for indexing activities?	yes	-	-	yes	-	yes	yes, efficiency

Question	IZA	Baldini	ALV	UUL	NBR	UOC	CDP
Do you need more time (%) to increase your expertise in usage (local systems)?	yes	-	-	yes	-	yes	yes
Do you need a stronger interaction between local system and remote librarian systems or general catalogues during the computer-based indexing?	yes	-	-	yes	-	yes	yes
Do you need more effectiveness in sw indexing systems?	-	-	-	yes	-	yes	yes
Do you need better on-line help and documentation for indexing?	yes	-	-	yes	-	yes	yes
On which technical areas do you need significant enhancement (detailed description)?	all	-	-	automatic indexing, event. Multi-lingual thesaurus	-	all	all
Which parts of your current indexing system (if any), would you like to have integrated in the Laurin system?	keywords, text-types	Reference to old keywords given by P. Monelli ...	-	-	-	working with thematic index (compatible thesauri)	working with thematic index (compatible thesauri)
Which special requirements (if any) should the Laurin system meet in the field of indexing?	easy handling, simple structure, online help	-	easy access through a thesaurus or a cross-ref.-list, eventually Full-text-search...	1) find some way to specify the importance/relevance in relation to the text of each keyword 2) find some way to index that articles belonging to the same topic can be linked together that satisfies the end-user	-	exact word research, boolean research, thematic research, etc.	exact word research, boolean research, thematic research, etc.

Index of pictures and illustrations