Renardus The Clever Route to Information

DDC Mapping Report

Renardus D7.4

Document details

Author:	Traugott Koch, Heike Neuroth, Michael Day
Date:	2001-02-15
Version:	0.1
Document Name:	d74-v01.doc
Notes:	Draft

Summary

This report provides background information on the classification mapping that is to be undertaken as part of the Renardus project. It contains general information on the benefits of Renardus-type services using subject classification schemes and some background information on the origins of the Renardus mapping effort. An assessment of the various Renardus partner gateways and their usage of classification schemes follows this, together with some background on the project's selection of the Dewey Decimal Classification (DDC) as the common classification system. The next few chapters deal with how to map subject relationships, some important issues, and the Renardus technical implementation. The final chapter deals with related work in the general area of mapping subject schemes.

NetLab

SUB

UKOLN

Author details

Michael Day, UKOLN: the UK Office for Library and Information Networking, University of Bath, Bath BA2 7AY, United Kingdom. m.day@ukoln.ac.uk

Traugott Koch, NetLab, Lund University Library Development Department, P.O. Box 134. S-221 00 Lund, Sweden and Department of Development & IT, Technical Knowledge Center & Library of Denmark (DTV), Technical University of Denmark (DTU), Anker Engelunds Vej 1, P.O. Box 777, DK- 2800 Lyngby, Denmark

Heike Neuroth, Niedersächsischen Staats- und Universitätsbibliothek Göttingen, Papendiek 14, D-37073 Göttingen, Germany. neuroth@mail.sub.uni-goettingen.de

Table of Contents

Glossary

AGRIS	A bibliographic database for agriculture - provided by the FAO
Agrovoc	A multilingual thesaurus of agricultural terminology
AIP	American Institute of Physics
ALA	American Library Association
ALCTS	Association for Library Collections & Technical Services
AMIA	American Medical Informatics Association
AMS	American Mathematical Society
BBK	Bibliothekarisch.Bibliographische Klassifikation
BC	Nederlandse Basisclassificatie (Dutch Basic Classification)
BC	Bliss Bibliographic Classification
BIOME	The RDN 'hub' for the health and life sciences
Biz/ed	A UK-based business and economics information service for students, teachers and lecturers
BK	Basisklassifikation (Basic Classification)
BMBF	Bundesministerium für Bildung und Forschung (German Federal Ministry for Education and Research)
BnF	Bibliothèque nationale de France
BUBL	Information service based at Strathclyde University, originally known as the Bulletin Board for Libraries
CARMEN	Content Analysis, Retrieval and MetaData: Effective Networking
CCM	Computing Classification System
CCS	ALCTS Cataloging and Classification Section
CENL	Conference of European National Librarians
CLRC	Central Laboratory of the Research Councils
CORC	Cooperative Online Resource Catalog
DAINet	Deutsche Agrarinformationsnetz (German Agricultural Information Network)
DBI	Deutsches Bibliotheksinstitut Berlin (now EDBI)
DDB	Die Deutsche Bibliothek

DDC	Dewey Decimal Classification
DDI	Data Documentation Initiative
DEF	Danmark's Elektroniske Forskningsbibliotek (Denmark's Electronic Research Library)
DESIRE	Development of a European Service for Information on Research and Education
DHSS	Department of Health and Social Security (UK)
DIPF	Deutsches Institut für Internationale Pädagogische Forschung (German Institute for International Educational Research)
DTD	Document Type Definition
DutchESS	Dutch Electronic Subject Service
DutchESS	Dutch Electronic Subject Service
ECOT	Educational courses and occupations thesaurus
EDBI	Ehemaliges Deutsches Bibliotheksinstitut
EELS	Engineering Electronic Library Sweden
EEVL	Edinburgh Engineering Virtual Library - the RDN 'hub' for engineering, mathematics and computing
Ei	Engineering Information
ERCIM	European Research Consortium for Informatics and Mathematics
ETH Zürich	Eidgenössische Technische Hochschule Zürich (Swiss Federal Institute of Technology Zurich)
FAO	Food and Agriculture Organisation of the United Nations
FVL	Finnish Virtual Library
GBV	Gemeinsamer Bibliotheksverbund der Länder Bremen, Hamburg, Mecklenberg-Vorpommern, Niedersachsen, Sachsen-Anhalt, Schleswig-Holstein und Thüringen
GERHARD	German Harvest Automated Retrieval and Directory
GHB	Gesamthochschulbibliotheken des Landes Nordrhein-Westfalen (University Libraries of Northrhine-Westfalia)
GOK	Göttinger Online-Klassifikation
HASSET	Humanities and Social Science Electronic Thesaurus
HEREIN project	European information network on cultural heritage policies

Humbul	The RDN 'hub' for the humanities
ICD9-CM	International Classification of Diseases: 9th revision, Clinical Modification
ICS-FORTH	Institute of Computer Science, Foundation for Research and Technology Hellas
IEE	Institution of Electrical Engineers
IEEE	Institute of Electrical and Electronics Engineers
IFLA	International Federation of Library Associations and Institutions
ILRT	Institute for Learning and Research Technology
INSPEC	A bibliographic database for physics, electronics and computing- provided by the IEE
ISKO	International Society for Knowledge Organization
IZ Sozialwissenschaften	Informationszentrum Sozialwissenschaften (Social Science Information Centre)
LCC	Library of Congress Classification
LCSH	Library of Congress Subject Headings
LIMBER	Language Independent Metadata Browsing of European Resources
MeSH	Medical Subject Headings
MPI Bildungsforschung	Max-Planck-Institut für Bildungsforschung (Max Planck Institute for Human Development)
MSC	Mathematics Subject Classification
NESSTAR	Networked Social Science Tools and Resources
NLM	National Library of Medicine
NOVAGate	Gateway of the
NSD	Norsk samfunnsvitenskapelig datatjeneste (Norwegian Social Science Data Services)
OASIS	Otlet's Adaptive Search Information Service
OCLC	OCLC Online Computer Library Center, Inc.
OMNI	Organising Medical Networked Information; a gateway that is part of the BIOME 'hub' of the RDN
PACS	Physics and Astronomy Classification Scheme

RAMEAU	Répertoire d'autorité-matière encyclopédique et alphabétique unifié
RDF	Resource Description Framework
RDN	Resource Discovery Network
RSWK	Regeln für den Schlagwortkatalog (Rules for the Subject Catalogue)
RVK	Regensburger Verbundklassifikation
SAC	ALCTS CCS Subject Analysis Committee
Sears	Sears List of Subject Headings
SIS	Semantic Index System
SIS-TMS	SIS Thesaurus Management System
SOSIG	Social Science Information Gateway
SQL	Structured Query Language
SSG-FI	Special Subject Guides - Fachinformation
SUB	Niedersächsischen Staats- und Universitätsbibliothek Göttingen
SWD	Schlagwortnormdatei (Subject Authority File)
TGN	Getty Thesaurus of Geographic Names
UDC	Universal Decimal Classification
UDK	Universale Dezimalklassifikation
UKOLN	UK Office for Library and Information Networking
UMLS	Unified Medical Language System
XML	Extensible Markup Language

1 Introduction

Quality-controlled subject gateways (sometimes called Internet information gateways) are services that provide access to selected Internet resources based on the creation of descriptive metadata. These services typically offer hierarchical browse structures based on subject classification systems (e.g., Koch, Day, *et al*., 1997). The gateways' use of classification schemes tends to be very pragmatic. Gateways want to provide their users with some kind of browse access to the catalogue of Internet resources that they provide and a hierarchical classification scheme fulfils this need quite well. The classification scheme is, therefore, being used for grouping in proximity those resources that share the same broad subject matter; in much the same way as classified books are organised in a systematic order on shelves in a physical library. Indeed, the emphasis in some gateways is not on producing a detailed classified catalogue of resources for improved information retrieval but to provide something more analogous to a relatively simple shelf location device. Guidelines for the use of classification schemes and other subject vocabularies by subject gateways can be found in a paper by Cross, *et al*. (2000).

Quality-controlled subject gateways use a wide variety of different schemes to give browse access to Internet resources. A list of Web-sites that is maintained by McKiernan (1996-2001) demonstrates that services (including subject gateways) that attempt to organise the Web by the use of classification schemes or controlled vocabularies use a wide variety of different standards. These include well-used (and known) general schemes like the Dewey Decimal Classification (DDC) and the Library of Congress Subject Headings (LCSH) as well as a number of more subject specialised or locally produced schemes, e.g., the Basisklassifikation (BK) of the GBV (Facharbeitsgruppe Sacherschließung, 2000).

Within the Renardus partner gateways, some services use general classification schemes; e.g. DutchESS uses the Dutch version of BK - the Nederlandse Basisclassificatie (BC); the SOSIG Internet catalogue uses a sub-set of the Universal Decimal Classification (UDC). However, other gateways use more specialised subject schemes; e.g. the EELS gateway uses the Ei classification codes developed by Engineering Information, Inc.

If the Renardus broker is required to give some kind of consistent browse access to records retrieved from more than one participant gateway, then it will itself need to utilise a single classification scheme. In order for this to work properly, it follows that participant gateways will either have to add the appropriate classification codes to their own records or to provide a mapping between the classification scheme used by the gateway to the scheme used by the Renardus broker. The premise of this report is that gateways will need to provide this mapping and to outline some guidelines for its completion.

The classification scheme chosen by the project to create browse access to the Renardus broker is the DDC. This is a well-established scheme, which was first published in 1876 by Melvil Dewey. It is currently in its 21st edition and is published by OCLC Forest Press (Mitchell 1995; Mitchell, et al. 1996). Research carried out by OCLC in the 1980s and 1990s established that the DDC is an appropriate tool for browsing, both for library catalogues and for the Internet (e.g., Markey 1989; Vizine-Goetz 1996a).

2 The benefit of classification in a service like Renardus

2.1 Introduction

Subject gateways typically give access to Internet resources by providing both searching and browsing facilities. The browsing functions of gateways, in particular, are usually dependent upon the adoption and use of some directory-like structure - often based on subject classification schemes or thesauri. The idea of classification is to make it easier for users to find and retrieve resources therefore browsing a directory-type structure is user friendly. The structure makes it relatively easy to navigate and is an important help if users are not looking

for a specific item. Users typically are able to choose categories from a subject hierarchy and to use these to make their way through the service, moving down the individual branches of a subject tree. A browsing structure also gives a helpful overview of the scope of a service and how a particular service is organised.

Subject classification is a method of describing resources by their subject and a means of organising knowledge in libraries and other information environments. Classification schemes differ from other subject indexing systems (subject headings, thesauri, etc.) by trying to create collections of related resources in a hierarchical structure. The use of notations facilitates the creation of hierarchical subject trees. By building a hierarchical structure, a classification scheme enables users to look for related items that have not previously been identified as relevant.

2.2 Advantages of using a classification system

Following Koch, Day, *et al*. (1997), a site that organises knowledge with a classification scheme has several distinct advantages over sites that do not:

- *Browsing*: Browsing is particularly helpful for inexperienced users or for users not familiar with a subject and its structure and terminology. The classification notation does not even need to be displayed on the screen so an inexperienced user can have the advantage of using a hierarchical scheme without the distraction of the notation itself.

- *Broadening and narrowing searches*: Questions can be limited to individual parts of a collection (filtering) and the number of false hits is reduced (i.e. for improved precision).

- *Context*: the use of a classification scheme gives context to the search terms used. For example, the problem of homonyms can be partly overcome.

- *Potential to permit multilingual access to a collection*: many classification schemes, especially the universal ones like DDC, are published in several different languages. In cases where notations independent from a specific language are used, these notations can serve as a switching language. Thus, such schemes can offer multilingual access to the same resources without the necessity to translate the description of a document.

- *The partitioning and manipulation of a database*: large classified lists can be divided logically into smaller parts if required.

- *Interoperability*: the use of an agreed classification scheme could enable improved browsing and subject searching across (distributed) databases.

- *Stability*: an established classification scheme is not usually in danger of obsolescence. The larger schemes are now undergoing continuous revision, although they are normally also formally published in numbered editions. Some classifications may have to be changed when a new edition of a scheme is published, but it is unlikely that every single resource will have to be re-classified.

- *Familiarity*: classification schemes have the potential to be well known. Regular users of libraries will be familiar with at least part of one or more of the traditional library schemes. Members of a subject community are likely to be familiar with their (subject-specific) schemes as well. Use of Internet services that use established classification schemes might therefore have an advantage over those that use a home-grown scheme or none.

- *Availability*: many classification schemes are available in machine-readable form.

Classification schemes, however, have some disadvantages:

The division of logical collections of material: classification schemes often split up collections of related material. This can be partly overcome with good cross-references or by the placing of documents into several classes. Furthermore some popular schemes do not always subdivide classes in a logical manner. This can make them difficult to use for browsing purposes. Also classification schemes often reveal difficulty in reacting to new areas of study and changing terminology since they are usually updated through formal processes by organised bodies Several of the general advantages as listed above apply to subject gateways on the Internet, especially. A classification system for example helps to bring together small collections of similar resources. This is of high interest in the Renardus context as well as that several distributed metadata collections can be searched with the same classification scheme.

The partners of Renardus agreed that all resources should be cross-browsable via a common classification system (cf. chapter 5). Besides other cross-searchable fields in the Renardus service like title, subject etc. the common browsing structure will be one highlight of Renardus. This is one main issue to ensure that users could have access to all high quality resources of partners' Subject Gateways through one important subject field, the classification system DDC (cf. chapter 5). The DDC allows the real subject access to distributed metadata collections.

3 The DDC mapping effort in Renardus: work description

The need for some kind of central browse structure based on a general subject classification scheme was noted by questionnaire responses for Renardus internal deliverable D6.4. This was discussed at the project meeting held in Paris on the 7-8 September 2000 and it was agreed that Traugott Koch (NetLab and DTV) would investigate the possibility of negotiating the use of the DDC from OCLC Forest Press. At a further project meeting, held in Göttingen on 6-7 November 2000, a small working group was set-up to begin formulating rules for the mappings and to oversee the mapping process. This group consisted of Traugott Koch (NetLab and DTV), Heike Neuroth (SUB), Michael Day (UKOLN) and Ewa Nieszkowska (BnF).

Three of this core group met in London on 18-19 December to discuss the development of some mapping guidelines, the production of a report and the content of a workshop for participating gateways. This workshop will be held in Göttingen on the 22-23 February 2001.

4 The partner gateways' usage of classification systems

4.1 Introduction

In the context of a second questionnaire developed and analysed within the framework of work package 6 (see Renardus internal deliverable D6.4 "data model and data flow") partners agreed that it is necessary to offer a common cross-browsing structure for Renardus. They also preferred DDC to be the common classification system and that all partners would need to map their classification system to DDC, and that this should be used to build up the browsing structure. Chapters 2 and 5 describe in more detail why a common browsing structure in Renardus is desirable and why the partners decided to use DDC.

This chapter will provide an overview of partners' classification systems and a preliminary analysis of their usage. Furthermore it will give some information on expected problems and/or of issues that will need to be considered. The last part provides some areas of subject overlapping between the Renardus Gateways.

4.2 Analysis of partners' classification systems

In Renardus, thirteen Subject Gateways from several different organisations, institutions or libraries will need to map their classification system to DDC. Some organisations will map their system to DDC on a broker level although they provide several Subject Gateways with

different classification systems (also) via a single platform (e.g. FVL, RDN with SOSIG, OMNI etc.).

The following table (Table 4.1) lists all of the participating Subject Gateways (with its responsible organisation and URL) that will take part in the Renardus mapping process.

Responsible Organisation	***Subject Gateway***	***URL Subject Gateway***
National Library of the Netherlands	**DutchESS**	http://www.kb.nl/dutchess/
Libraries of the virtual NOVA University in Denmark, Finland, Iceland, Norway and Sweden (Libraries of Nordic agricultural and veterinary universities)	**NOVAGate**	http://novagate.nova-university.org/
NetLab and the Swedish Universities of Technology Libraries	**EELS**	http://eels.lub.lu.se/
Danmarks Elektroniske Forskningsbibliotek	**DEF fagportal**	http://www.deff.dk/vejviser/index.zap?sprog=eng
Zentralstelle für Agrardokumentation und -information, Germany (ZADI)	**DAINet**	http://www.dainet.de/dain-eng/suche/index.htm
Jyväskylä University Library, Finland (JyU)	**FVL**	http://www.jyu.fi/library/virtuaalikirjasto/engvirli.htm
National Library of France (BNF)	**Les Signets**	http://www.bnf.fr/web-bnf/liens/
DNER Office, King's College London, United Kingdom	**RDN**	http://www.rdn.ac.uk/
State Library of Lower Saxony and University Library of Göttingen, Germany (SUB), SSG-FI Guides	**Anglistik Guide**	http://www.AnglistikGuide.de/
see above	**Geo-Guide**	http://www.Geo-Guide.de/
see above	**History Guide**	http://www.HistoryGuide.de/
see above	**MathGuide**	http://www.MathGuide.de/
National Library of Germany (DDB)	**Deposit Server**	http://deposit.ddb.de/index.htm

Table 4.1: Overview of Subject Gateways participating in the first mapping

Table 4.2 provides a general overview of partners' classification systems. Most of the Subject Gateways use a thematic classification system, only 5 of 13 gateways support a universal system. Only one Subject Gateway has no classification system, but subject headings for a browsing structure sorted in an alphabetically list (DAINet, see below). The majority of

classification systems are more or less local or national. International systems are mainly based on the Basisclassificatie (BC) or Dewey Decimal Classification (DDC) or adaptations of them. Two classification systems seem to be more general than DDC: BC (DutchESS, see table 3) and SWD (Deposit Server, see table 3). Since January 2001 the DDB (Deposit Server) will catalogue the online dissertations also directly with DDC. BnF (Les Signets) will use DDC in more than one level in the future, the library is re-organising their DDC based classification system.

Nearly all subject gateways offer an English classification language and thus an English browsing structure. Only the Deposit Server (furthermore SWD is only in German) and DAINet offer no hierarchically structured browsing page. Until now it is not clear whether and how the DDB will realise a browsing entry for their online theses. DAINet will offer a browsing structure in the near future (see table 4). Now browsing is realized through alphabetically lists for subject headings of SUBJECT and OBJECT thesauri (see below).

Searching in / via notation is not possible for all subject gateways (e.g. DutchESS, NOVAGate). Whether this search option will be a recommendation for the Renardus service is not quite clear yet.

Subject Gateway	***thematic, universal***	***international, national, local***	***classification language(s) in use***	***browseable***	***searchable (via notation/caption)***
DutchESS	universal	international (BK adaptation)	Dutch, English	yes	no for notation, caption?
NOVAGate	thematic	international	English	yes	no for notation, caption?
EELS	thematic	international	English	yes	yes (notation and caption)
DEF fagportal	universal	local (DDC based)	Danish, English	yes	searchable both separately and jointly
DAINet	home-grown: thematic (subject headings)	international	German, English	**no**	no for notation, in future yes via subject headings
FVL	home-grown: thematic	local	Finnish, English	yes	yes via descriptor
Les Signets	DDC based universal	international (DDC adaption)	French	yes	no
RDN hubs	home-grown: universal	local	English	yes	no
Anglistik Guide	thematic	local	English	yes	yes (both)
Geo-Guide	thematic	local	English	yes	yes (both)
History Guide	thematic	local	English	yes	yes (both)
MathGuide	thematic	local	English	yes	yes (both)

Subject Gateway	*thematic, universal*	*international, national, local*	*classification language(s) in use*	*browseable*	*searchable (via notation/caption)*
Deposit Server	home-grown: universal	national	German	**no**	no for captions, yes for notations

Table 4.2: General overview of the classification systems used by partners

Table 4.3 offers an overview of the relevant classification systems for the real mapping work to DDC and further information about the number of hierarchical steps (level), number of classes per level and total number of classes that have to be mapped to DDC. Sometimes the total number of the official classification system used by a partner and the number of classes that have to be mapped to DDC is different. For example DutchESS uses BK, but not all BK notations and captions are in usage. MathGuide offers in the browsing page two or three levels but catalogues in much more detail than is shown in the MathGuide browsing structure. The question is how deep the mapping from partners' classification systems to DDC has to be. It is not imaginable that in the Renardus context the whole BK or MSC could map to DDC.

Most of the used thematic systems are very special and in a deep structure like EELS or all of SSG-FI gateways. EELS (Ei) and MathGuide (MSC) use an international thematic classification system. DAINet offers the browsing structure with a detailed and hierarchical list of subject headings, which is based on two different thesauri. All other thematic systems (NOVAGate and FVL) are not so extensive: one or two levels maximum and in total 18 and 60 numbers of classes have to be mapped to DDC. For EELS about 800 thematic classes have to be mapped, which are structured in 5 levels. For SSG-FI Guides between 200 and 400 classes, structured hierarchically in up to 5 levels have to be mapped.

Up to now it is not clear how RDN will realise the mapping work. Several classification systems are used in the different hubs. For example OMNI uses NLM and LCC, AgriFor uses DDC and EEVL uses a home-grown system (loosely) based on Ei classification (see Renardus D6.1 Appendix F: Overview Classification Systems at http://renardus.sub.uni-goettingen.de/wp6/d6.1/index.html). It makes sense to map first on RDN broker level with hierarchically structured subject headings and later on, after some experiences with the real mapping work, to map on Subject Gateway level.

The SSG-FI Gateway History Guide will probably change the primary classification system from GOK to DDC in the context of co-operation with another German history Gateway. For all Subject Gateways that use primarily the DDC it has to be checked whether this DDC notation could directly be re-used for Renardus purposes or whether there has to be developed a special mapping from the official DDC to the Renardus specific DDC. For example the differences between the DDC adaptation used for the Gateway Les Signets and the "official" DDC is not clear yet. Maybe these differences arise only because of the French translation of the English DDC version published by OCLC Forest Press.

Subject Gateway	*relevant classification system for mapping*	*levels*	*number of classes/level*	*total number of classes*
DutchESS	BK adaption	2	main classes: about 4, second classes: 48, third classes: about 800 - 1000 (??)	?

Subject Gateway	*relevant classification system for mapping*	*levels*	*number of classes/level*	*total number of classes*
NOVAGate	FAO/AGRIS main subject categories	1	main classes: 18	18
EELS	Ei classification system	5	main classes: 6, second classes: 38, third class: 182, etc.	800
DEF fagportal	top level shallow classification (based on DDC)	2	main classes: 4, second classes: 26	30
DAINet	no classification, home-grown: subject headings (based on two thesauri: SUBJECT and OBJECT)	SUBJECT: 2 OBJECT: 2	SUBJECT: main classes: 9 second classes: 100 OBJECT: main classes: 10 second classes: 80	SUBJECT 109 OBJECT 90
FVL	home-grown: main subject headings	2	main classes: 4, second classes: about 56	60
Les Signets	DDC adaptation	1	??	??
RDN hubs	home-grown: main subject headings	2 (??)	main classes: 9 (soon 11), second level classes: about 45 (??)	about 60
Anglistik Guide	GOK	5	main classes: 5, second classes: 5, third classes: 42, fourth classes: 87, fifth classes: 52	190
Geo-Guide	GOK	4	main classes: 18, second classes: 131, third classes: 150,	317

Subject Gateway	*relevant classification system for mapping*	*levels*	*number of classes/level*	*total number of classes*
			fourth classes: 18	
History Guide	GOK	5	main classes: 12, second classes: 83, third classes: 68, fourth classes: 198, fifth classes: 18, (1 sixth)	380
MathGuide	MSC	3 in MSC 2000	general classes (MathGuide specific): 7 main classes: 65, second classes: 420, third classes: 2100	to map probably up to the second MSC 200 level: about 500
Deposit Server	home-grown shallow classification: SWD	4	main classes: 37, second classes: 186, third classes: 236, fourth classes: 21	480

Table 4.3: Overview of the structure of classification systems (level, number of classes etc.) used by partners

Table 4 references partners' browsing structures realised via a classification system or subject headings and mentions some general notes or comments regarding the classification system or subject headings which will be mapped to DDC.

Subject Gateway	*URL of browsing page*	*notes/comments*
DutchESS	http://www.kb.nl/dutchess/nbc_main.html	A version of BK is in use: in official BK: about 50 main classes and about 2000 second classes
NOVAGate	http://novagate.nova-university.org/browse.html	
EELS	http://eels.lub.lu.se/ (6 browsing entry points)	
DEF fagportal	http://www.deff.dk/vejviser/index.zap?sprog=eng (4 browsing entry points)	

Subject Gateway	*URL of browsing page*	*notes/comments*
DAINet	not available (planned for future)	DAINET will map up to 2 levels although both thesauri consist of 7 (SUBJECT) resp. 5 (OBJECT) levels with 500 resp. 600 total number of classes
FVL	http://www.jyu.fi/library/virtuaalikirjasto/engvirli.htm	FVL will map on broker level
Les Signets	http://www.bnf.fr/web-bnf/liens/	BnF is re-organising the classification of Les Signets. It would have, at least, 3 or 4 clear levels. At the moment only the first level can be shown
RDN hubs	http://www.rdn.ac.uk/	RDN will probably first map on Broker level
Anglistik Guide	http://www.anglistikguide.de/cgi-bin/ssgfi/navigator.pl?db=lit&type=subject/subj	The GOK version used here slightly deviates from the official GOK system
Geo-Guide	http://www.Geo-Guide.de/cgi-bin/ssgfi/navigator.pl?db=geo&type=subj	The GOK version used here slightly deviates from the official GOK system
History Guide	http://www.HistoryGuide.de/cgi-bin/ssgfi/navigator.pl?db=hist&type=subj	The GOK version used here slightly deviates from the official GOK system
MathGuide	http://www.MathGuide.de/cgi-bin/ssgfi/navigator.pl?db=math&type=subj	Catalogued more detailed, but for browsing structure only 2 levels are displayed. MathGuide introduces a more general level (7 classes) for structural reasons which is not foreseen in MSC 2000.
Deposit Server	not available	??

Table 4.4: References and general notes of classification systems used by partners

Two Subject Gateways offer until now no hierarchically structured browsing page (see table 4):

DDB:

Esther Scheven sent the following information regarding their classification system: "The online-theses have a keyword string and this keyword-string is classified by the SWD-Classification. That means that each string may consist of several SWD-notations like the keywords. But the notation of the keyword string is not a enumeration of the notations of the keywords. It is even possible that the notations of the keyword string is different from the individual keywords". The SWD classification system is not available in an online version thus users can't browse along the SWD to browse for online thesis. Only searching via notations or keywords is possible (see `http://z3950gw.dbf.ddb.de/z3950/zfo_get_file.cgi?fileName=DDB/searchForm.html`). Furthermore the whole interface and the classification system are only available in German language.

Example of hierarchical structure of SWD:

1. Level: **10** Wirtschaft, Verkehr, Umweltschutz, Raumordnung

    2. Level: **10.1** Wirtschaftsgeschichte

    2. Level: **10.2** Wirtschaft, Volkswirtschaft

    3. Level: **10.2a** Allgemeines

        4. Level: **10.2aa** Volkswirtschaft

        4. Level: **10.2ab** Wirtschaftssystem

        4. Level: **10.2ac** Mathematische Methoden, Information, Entscheidung

    3. Level: **10.2b** Haushalt, Verbraucher

    3. Level: **10.2c** Mikroökonomie, Wettbewerb

    3. Level: **10.2d** Konjunktur, Verteilung, Wirtschaftsstruktur

        4. Level: **10.2da** Wirtschaftskreislauf, Konjunktur

        4. Level: **10.2db** Verteilung

        4. Level: **10.2dc** Wirtschaftsstruktur

It has to been checked how the Renardus users can jump from the DDC browsing pages to the DDB browsing structure. But the SWD as a general classification system is not so detailed (only about 480 classes in total) as the DDC classification system so users will not get more browsing information when jumping to the local Subject Gateway browsing pages.

DAINet:

DAINet uses no classification system but subject headings generated from two home-grown thesauri. These thesauri are a SUBJECT and an OBJECT thesaurus, based on the internationally known thematic AGROVOC thesaurus. The AGROVOC thesaurus is built polyhierarchically, which means that subject headings from the OBJECT thesaurus are classified several times under the SUBJECT thesaurus and that forms one AGROVOC thesaurus. DAINet thesauri are structured monohierarchically, so DAINet needs two separate thesauri and thus two separate browsing pages: one for browsing after SUBJECT (see http://www.dainet.de/dain-eng/suche/index.htm under Scientific Topics) and one for browsing after OBJECT (see http://www.dainet.de/dain-eng/suche/index.htm under Products/Factors). Furthermore, the browsing is only possible via an alphabetical list of SUBJECT or OBJECT subject headings, no hierarchically structured subject- or object-related browsing is possible.

For the future DAInet plans to offer a hierarchical browsing structure. So far it is not clear if DAINet brings together the two thesauri like AGROVOC had done and if DAINet introduces notations for the subject headings based on the AGRIS classification system. It seems not so difficult to reuse the AGRIS system, because this system is based on the AGROVOC thesaurus and both relevant DAINet thesauri are also based on AGROVOC.

Marianne Andres from DAINet sends some first impressions regarding the mapping work to DDC: "Most of the special terms of the DAINet thesauri have no appropriate term within the DDC. It is necessary to map the special DAINet term to a more general term of the DDC. Because of the quite other building of classes within the DDC there will arise other relationships when descriptors of the DAINet thesauri are mapped to relevant DDC terms. Particularly this will be fact in those cases, if agricultural related terms of DAINet thesauri must be mapped to quite other classes of DDC not belonging to the chapter 630 (Agriculture),

i.e. DAINet subject fields economy or environment. For a far-reaching complete mapping it is necessary to bring at least two different DAINet thesauri into accordance with DDC".

4.3 Subject overlap between Renardus Gateways

The gateways have a wide collection scope, with the following core subjects:

- DutchESS aims to cover all disciplines. Since most of the participating libraries have been humanity libraries, currently humanities and social sciences are covered best
- NOVAGate's core subjects are forestry, veterinary, agricultural food and environmental sciences
- EELS covers the subjects engineering and related sciences, polar region research
- DEF fagportal has the main collection focus on energy, food science, clinical medicine, music, and industry economics
- DAINet covers all disciplines in agriculture, nutrition, forestry, and fishery
- FVL has 55 subjects within the fields of culture, language & education, society and economy, health and welfare, and natural sciences and technology
- Les Signets have a general collection scope with about 130 different subjects
- RDN's current core subjects are health, life sciences, engineering, mathematics, computing, social sciences, business, law and the humanities. Gateways to cover other subjects will be added in the near future.
- SSG-FI with its four guides has the following core subjects:

  Anglistik Guide: British and American literature
  Geo-Guide: earth sciences, geography, thematic maps, and mining
  History Guide: Anglo-American history
  MathGuide: mathematics

A comparison of the core subjects of all gateways shows several overlaps that could be of interest for the common classification mapping work. Here are examples for congruencies in subject fields:

- **engineering**: EELS, RDN (EEVL), DEF fagportal in the field of energy, technology, NOVAGate, Geo-Guide for applied sciences
- **humanities**: DutchESS, RDN (Humbul), SSG-FI (Anglistik Guide, History Guide), FVL
- **forestry and agricultural (and food) sciences**: NOVAGate, DAINet, EELS, FVL, DEF fagportal (technology and sciences: food sciences), RDN (AgriFor), partly Geo-Guide and soon SSG-FI ForestryGuide
- **mathematics**: SSG-FI (MathGuide), RDN (MathGate), EELS, EEVL
- **economics, social sciences**: RDN (SOSIG, Biz/ed), FVL, DEF fagportal (social sciences)
- **medicine**: RDN (OMNI), DEF fagportal (technology and sciences: medicine, health and nutrition)
- **earth sciences, geography**: DEF fagportal (technology and sciences: geology. meteorology. palaeontology. geography and travel), Geo-Guide, partly EEVL, EELS

5 The reasons for the selection of DDC as common classification system for browsing

5.1 Introduction

The Renardus service will give access to resources from all areas of subjects, published world-wide and in many languages (see Evaluation report of partner subject gateways, D6.1 public Renardus report: http://renardus.sub.uni-goettingen.de/wp6/d6.1/html/d6_1/doc0002.htm) and is intended to be offered to an international multi-disciplinary community of users (see Scoping Document , D6.2 public Renardus report http://www.renardus.org/deliverables/d6_2/D6_2summ.html).

In the context of the internal deliverable D6.4 ("data model") a second questionnaire was developed by SUB. Among other things partners were asked to evaluate several questions regarding classification aspects (required, strongly recommended, recommended, desirable, not necessary, and definitely not). The following list provides the most important statements:

It is required or strongly recommended that:

- Subject Gateways should be browsable via a common classification system,
- Renardus should use an existing common classification system,
- the common classification system should be provided in several European languages.

Considering these issues an existing universal classification system should be selected to build the common browsing structure in Renardus. DDC and UDC have a good multilingual capability due to the fact that the codes they produce are entirely numerical and their schedules have been widely translated (into up to 30 different languages). Furthermore universal classification systems like DDC and UDC are used by many Internet services and are readily available in machine-readable form. Using such well-known and international distributed classification systems guarantees maintenance by the owner of this classification system. For a service like Renardus with several distributed special subject gateways involved interoperability issues are also very important. Some gateways already use DDC or provide a mapping to DDC. An analysis of the classification systems used by the Renardus partners will also be provided in this report (see chapter 4). In the second questionnaire (D6.4) partners agreed that the common classification system should be DDC.

5.2 The main arguments for using the DDC

The DDC was finally preferred to the UDC as the classification system for cross-browsing in Renardus for the following reasons:

5.2.1 Online availability and tools:

1.1 The DDC has been much earlier and faster adapted to the usage in digital systems via the net. Apart from "DDC for Windows" the full classification incl. all support information is easily available as "Web Dewey" for all Web browsers and platforms.

The UDC Master Reference File however is only delivered as a file in two different file formats. Even when only to be used as a classification tool, a special administration system and user interface needs to be developed and maintained.

1.2 The DDC is via the Web connected to large databases of real documents that allow the verification of the potential subject content of classes and the correct and common usage of the classification. Apart from the traditional library catalogues, World Cat and Library of

Congress Catalogue with tens of millions of books etc., many other online available catalogues on the Web can be used in this way.

With CORC is the largest database of catalogued and classified Internet resources available in an integrated way, searchable and browsable with the enhanced DDC.

In this respect, both as an important classification and mapping support and as necessary prerequisite for cross-searching of heterogeneous databases on the Internet, the UDC falls far behind.

5.2.2 Global usage:

2.1 The DDC has a much larger user community on a global level and even the UDC's former advantage in Europe has more or less been lost.

Tens of thousands of libraries are using the DDC and so do 56 national bibliographies. The UDC's main users are special libraries and libraries in Eastern European countries. Important users have been moving away from the UDC in the last couple of years.

2.2 After an initial good position for the UDC as classification system in Internet resource discovery services the DDC is now more frequently used. For example, a well-known service like BUBL has abandoned the UDC for DDC. (cf. Koch, Day, *et al*. 1997)

5.2.3 Suitability of the classification system and its functionality:

The UDC is probably more "modern" and has made faster efforts towards a facetted structure. Its way of allowing the construction of new special classes by a straightforward relationship between existing ones is superior to the approach of build numbers in the DDC. In the Renardus application context however the way this is normally used turns into a disadvantage. Especially the special libraries using the UDC try to construct very special and often local classes (sometimes used to replace detailed keyword indexing) which are not shared by many other institutions world-wide. The UDC is, in addition, with 61 000 classes three times larger than the DDC and might be far too detailed for the Renardus purpose.

The outcome is a reduced interoperability between different applications and thus a reduced usability in a cross-browsing context like in Renardus. Such a service that maps between existing classification structures in heterogeneous services needs as many globally used and well defined classes as possible and has to avoid all the rich flexibility.

5.2.4 Frequency and character of the updates:

Both systems have large amounts of outdated captions and class structures and have understandable problems to keep up with the changes in vocabulary and scientific development. DDC seems to have a slight advantage in the speed and frequency of updating of main classes. Auxiliary tables where UDC is strong are not that important to the Renardus application. The UDC shows - for reasons that can be understood - much less US bias in its vocabulary and structures. Both are slow with integration of external standard vocabularies.

The main advantage of DDC here is the effort to adapt the captions of the system to general end-users and, even more important to Renardus, to the real content of digital documents in today's Internet. The necessary background for this is years of work with Internet discovery services using the system for classification that OCLC did and does with the DDC (CORC and predecessors).

5.2.5 Research and methodological development efforts:

With the rich, broad and long-term activities of the OCLC Office of Research Knowledge Organization Group and related Dublin Core metadata and XML/RDF work, the DDC has a

decisive advantage when it comes to research based development potential and future prove in digital library contexts.

So contains the enhanced DDC Renardus intends to use intellectually and statistically mapped vocabularies like the LCSH which are extremely useful in the classification and mapping work as they relate up-to-date vocabulary and subject terminology from other databases to the individual DDC classes. This is a necessary step to prepare options of automatic and semi-automatic classification efforts that are so important in the context of digital libraries.

In many important areas Renardus partners, e.g. in the DESIRE project (Koch and Vizine-Goetz, 1998) already have years of co-operation with OCLC (as the owner of the DDC): in metadata and Dublin Core development and implementation, in knowledge organisation and mapping methodologies, in encoding and syntax standardisation work like RDF and XML, in the area of automatic classification and when it comes to develop the preconditions for distributed usage of vocabulary systems.

In summary, UDC might be a good alternative to DDC in traditional classification usage. When it comes to digital library applications however and especially to the classification mapping task in Renardus, the UDC system and its development efforts are clearly insufficient and fall far behind the DDC.

If we like this situation or not, the DDC is today the only realistic alternative for Renardus and related projects.

5.3 The DDC research license

The basis for the usage of the DDC in the Renardus project is a research agreement with OCLC Forest Press, the owner of the DDC.

It allows Renardus to use the full enhanced DDC classification system (as used and displayed in the CORC service, with four updates a year) to construct and offer common Renardus cross-browsing pages during the lifetime of the Renardus project.

DDC notations shall not be displayed.

Renardus can adapt captions of DDC classes to European vocabulary as long as they express the same coverage. Changes for other reasons are not expressively allowed.

According to the agreement only the English language version of the DDC can be used in the Renardus browsing pages. A future aim for Renardus is to use several European languages as far and deep as available.

OCLC Forest Press gets an irrevocable right to explore the Renardus classification mappings.

Future terms and conditions for the time when the project is finished but Renardus hopefully is continued as a stable public service are not known yet. They will be discussed before September 2001. Having invested all this effort into the classification mapping, the partners are eager to preserve conditions which allow to continue with terms that allow our institutions to offer Renardus as a sustainable public Internet service.

The text of the agreement is available to Renardus partners on the internal Web site.

In addition to support with technical aspects of the usage of DDC the OCLC Office of Research Knowledge Organization Research Group will co-operate with the Renardus mapping group in questions regarding methodological and theoretical development.

6 Mapping relationships: why and how. Alternative mapping languages

6.1 Why? The differences are crucial for a service like Renardus

Many traditional and less advanced mapping projects like conversions or concordances between pairs of two classification systems for usage in OPACS or union catalogues limit themselves to just establishing a connection between pairs of classes. They leave it open what the character and degree of the indicated equivalence is.

The cross-browsing service in Renardus aims to mediate between many different and heterogeneous classification systems using the DDC classification as a common "switching language" and browsing structure. The structure and level of detail, the vocabulary, language and cultural context is extremely different between these locally applied classification systems and the universal DDC.

Therefore we expect a straight and full equivalence between the content of two classes to be a rather rare situation. The same judgement has been made by other related projects like CARMEN (cf. below).

In the Renardus Subject browsing pages the user needs to be told that certain links from a DDC class point to a class in a local gateway containing a broader or narrower area of content, showing major or minor overlap with the DDC class. That is especially true since there quite often will be mapping links to several different classes in different subject gateways. One might be fully equivalent, another only showing a minor overlap.

The need for a more detailed specification of the degree of equivalence is even greater when the mapping between the local class and the DDC classes is used in the Renardus advanced (subject) searching feature. Records from a class with only a minor overlap to the DDC class searched for or hit by the search should not be displayed in the same way and position as hits from a class fully equivalent to this DDC class. The ranking and sorting algorithms of the retrieval system need to be influenced by these differences in equivalence and mapping relationships.

Finally there are some reasons for this extra effort of specifying the mapping relationships emanating from a perspective beyond the Renardus project. From the literature and projects in the area of ontologies, knowledge engineering and knowledge management, it seems clear that advanced knowledge engineering applications need a much higher (and more costly) level of semantic relationship analysis, more advanced indexing and classification methods and systems than used in the library world so far, in order to explore the full richness of large collections of digital full-text documents and their catalogue surrogates/metadata. Future advanced and value-adding services might want to use the Renardus subject gateway data in this way. A database created by harvesting the full-text of documents catalogued in Renardus subject gateways, adding related documents and offering all that for integrated searching could be one such application.

6.2 How? The mapping process and language

6.2.1 Mapping process

A few practical principles are needed to keep the mapping work consistent and the resulting Renardus browsing pages balanced. They might, as well as the format and the examples, become slightly changed and amended during the real mapping work that is still ahead of us.

Every record in the mapping information file expresses the relationship between one class in the DDC and one class in the local classification system. If the same class has to be mapped to other classes this requires separate mapping relationship records.

When treating a certain class one should look at the same level of specificity of the subject content and try to find a fully equivalent class first, then look for true subset or superset classes and finally watch for overlapping situations (cf. the mapping language below).

If there is no similar specificity level available in the local classification the mapping should aim at the next higher level in the local hierarchy.

The mapping process, as reflected on the Renardus browsing pages, should be completely finished for the top level of the hierarchy and then move downwards in the local hiererchy, to assure a balanced Renardus service at all times. The final goal is, of course, to map all local classes to the DDC. This general rule does not prevent from starting the mapping work at the deepest levels in one branch and working the way up to the top level, which is often a better approach. Another suitable approach is to identify fully or good equivalence classes first and go up and down from there.

6.2.2 Mapping information format

The format is as follows (cf. 8.4 and the mapping guidelines for further details):

LN Local classification notation
LC Local classification caption
RL Relationship code
DN DDC notation
DC DDC caption
UR Local browsing URL for the class
NT Notes
:RE: Record end

In case we manage to adapt the CARMENx mapping tool to the Renardus mapping effort this format will be produced by the tool from the entries made into it.

A certain amount of practical experience with the mapping and the resulting browse and search features might reveal that we are able to omit one or two of the relationships in our practical application.

6.2.3 Relationship codes, definitions and rules

The following are the mapping relationships we foresee to use:

They are chosen to be as close to the solutions of related projects (as described below) as possible considering the special Renardus application.

The local class is:

- Fully equivalent. Code =
- Narrower equivalent. Code
- Broader equivalent. Code
- Major overlap. Code MJ
- Minor overlap. Code MN

As opposed to other projects we do uni-directional mapping only, from the DDC classification to the local classification(s). This explains the orientation and naming of the symbols above,

e.g. a certain DDC class a local class means that the local class is a narrower equivalent to the DDC class.

Cf. the `illustrations`.

The mapping relationships express the relationship between the subject contents intended to be placed in the DDC class and the local class respectively. Only one relationship can be assigned to a certain pair of classes.

The three types of equivalence require that one of the two classes is a true subset of the other, that it is not to be mapped to another class of the comprising classification. Full equivalence is the intermediate situation where both classes are basically 100% equivalent.

The two overlapping relationships require that parts of both classes do clearly not belong to the subject content of the other class.

1. **Fully equivalent** means more or less fully, not necessarily totally congruent. Fully equivalent should be the only mapping relationship for both classes.

Example:

Fully equivalent:
LN: 400
LC: Civil Engineering
RL: =
DN: 624
DC: Civil engineering
UR: http://eels.lub.lu.se/ei/40.html
NT: discuss this with Martin
:RE:

The remaining relationships express either the situation that one of the two classes mapped to each other is basically fully comprised in the other (narrower and broader equivalent) or the situation that both classes are only partly overlapping each other (major and minor overlap). In all these cases one or both classes will be involved in other mapping relationships too.

2. (The local class is) a **narrower equivalent** means that basically the potential content of the whole local class is comprised in the DDC class. The local class is a true subset of the content of the DDC class.

Narrower equivalent:
LN: 441.1
LC: Dams
RL:
DN: 627.8
DC: Dams and reservoirs
UR: http://eels.lub.lu.se/ei/441.1.html
NT: alt.: map to most subclasses of 627.8
:RE:

3. Vice versa, (the local class is) a **broader equivalent** means that basically the potential content of the whole DDC class is comprised in the local class. The local class is a true superset of the DDC class.

Broader equivalences from two DDC classes to the same local class:
LN: 406.2
LC: Roads and streets
RL: <

DN: 625.7
DC: Roads
UR: http://eels.lub.lu.se/ei/406.2.html
NT:
:RE:
LN: 406.2
LC: Roads and streets
RL: <
DN: 625.8
DC: Artificial road surfaces
UR: http://eels.lub.lu.se/ei/406.2.html
NT:
:RE:

Broader equivalences from three DDC classes to the same local class:
LN: 411
LC: Bituminous Materials
RL: <
DN: 691.96
DC: Bituminous materials
UR: http://eels.lub.lu.se/ei/411.html
NT: 411 is under Construction materials; 691.96 is under building materials
:RE:
LN: 411
LC: Bituminous Materials
RL: <
DN: 625.85
DC: Bituminous materials
UR: http://eels.lub.lu.se/ei/411.html
NT: 625.85 is Bit. mats. in pavements
:RE:
LN: 411
LC: Bituminous Materials
RL: <
DN: Asphalt concrete--sidewalk engineering
DC: 625.885
UR: http://eels.lub.lu.se/ei/411.html
NT: 625.885 is in sidewalks, RI term
:RE:

The degree of **overlap** is seen from the perspective of the DDC class. The mapping points to a local class.

4. If the local class has a major overlap with the subject content of the DDC class, related to the full content of the DDC class, the relationship is MJ: Major Overlap.

Major overlap:
The local class has a major overlap with three different DDC classes
LN: 411.1
LC: Asphalt
RL: MJ
DN: 691.96
DC: Bituminous materials
UR: http://eels.lub.lu.se/ei/411.1.html
NT:
:RE:411 is fully equiv. to three DDC classes
LN: 411.1
LC: Asphalt

RL: MJ
DN: 625.85
DC: Bituminous materials
UR: http://eels.lub.lu.se/ei/411.1.html
NT:
:RE:
LN: 411.1
LC: Asphalt
RL: MJ
DN: Asphalt concrete--sidewalk engineering
DC: 625.885
UR: http://eels.lub.lu.se/ei/411.1.html
NT:
:RE:

5. Consequently, if the intended local class content covers less than half of the intended subject content of the DDC class, it is called a minor overlap. The local class will then be mapped to several DDC classes. This is expected to be the most frequently occurring situation for most Renardus gateways.

Minor overlap:
LN: 440
LC: Water and Waterworks Engineering
RL: MN
DN: 627
DC: Hydraulic engineering
UR: http://eels.lub.lu.se/ei/440.html
NT: close to 50% overlap
:RE:

The number of resources found in the local class at a given point in time is of no relevance to this relationship and overlap judgement. E.g. if the intended content of the local class covers most of the content of the DDC class we have a major overlap. This is not changed just because the real resources in this class in the gateway might belong to only one small aspect of the intended subject content of the local class. The main reason for this rule is that the resource content of a certain class in a gateway can and will change over time. We are mapping classes to each other and not individual resources.

More difficult examples, i.e. combinations of the above-mentioned relationships emanating from the same set of classes, will be illustrated later.

6.2.4 Outcome on the Renardus browsing pages (notations will be omitted)

Emanating from some of the examples above:
(691) Building materials
(691.96) Bituminous materials
Broader: EELS (411) `Bituminous materials`
Major overlap with: EELS (411.1) `Asphalt`

(627) Hydraulic engineering
(627.8) Dams and reservoirs
Narrower:
EELS (411.1) `Dams`
EELS (411.2) `Reservoirs`

(627) Hydraulic engineering
Minor overlap with: EELS (440) `Water and Waterworks Engineering`

6.3 Alternative mapping languages

We have investigated several mapping languages in use or proposed for real mapping projects before deciding on our own solution. The reasons are obvious: a) not to oversee relationships similar projects are finding important and b) to stay as close as possible to other solutions in order to facilitate a certain level of interoperability between the mappings. One could imagine a valuable increase of the mapping universe if other projects or classification owners would map their systems to the DDC in a similar way and with a mapping language close to the one used by Renardus. Other useful chains of bilateral mappings would occur when anyone of the local Renardus classifications is mapped to a different third classification system. Mappings via more than one intermediate classification are however most probably not correct or useful anymore.

We had a closer look at the following mapping languages:

- Doerr 2001
- CARMEN
- Iyer and Giguere 1995
- CARMENx

The first two solutions relate to thesaurus mapping, Iyer and CARMENx to classification mapping.

6.3.1 Doerr 2001

Doerr (2001) proposes a slight extension of the ISO 5964-1985 [Guidelines for the establishment and development of multilingual thesauri] standards equivalence expressions by interpreting them as set relations of the associated sets. This provides the expressive power of Boolean queries.

- "partial equivalence" should become "broader equivalence" (is subset of) or "narrower equivalence" (is superset of)
- "exact equivalence" is interpreted as "same set as"
- "inexact equivalence" is interpreted as "overlaps with"
- "single to multiple equivalence" should become "*equivalence" to "compound" where "compound" is a Boolean expression of target terms with AND, OR, NOT and "*" is either "exact" or "broader" or "narrower".

According to Doerr, solutions chosen by the Getty Information Institute 1996 [Guidelines for forming language equivalents: A model based on the Art and Architecture Thesaurus], the on-going HEREIN project [http://www.european-heritage.net/fr/Thesaurus/Contenu.html] and the UMLS Metathesaurus (National Library of Medicine, 2001) could be subsumed under this proposal.

6.3.2 CARMEN

The types of relations used in the thesaurus mapping parts of the CARMEN project as presented by B. Woldering (of DDB) at a Renardus meeting in Göttingen, Nov 6, 2000:

- Equivalents:

- congruent descriptors ("=")
- equivalent descriptors ("=")
- AND-combinations ("=+ / +=")
- OR-combinations ("o= / =o")
- AND-OR-combinations ("=+o / o+=")
- double OR-combinations ("o=o")
- Broader term, as single term or in combination ("<")
- Narrower term, as single term or in combination (">")
- Related term, as single term or in combination ("^")
- No relation possible ("0")

It is not known at the moment, which of these relations are really used in the project. CARMEN is using Doerr's SIS-TMS software [SIS-TMS] as mapping support and hence, the mapping languages seem very similar.

6.3.3 Iyer and Giguere 1995

Iyer and Giguere (1995) proposed a mapping language for a planned prototype expert system interface to map between the mathematical classification MSC and the DDC20. As far as we know, this mapping has never been carried out.

They identified the following mapping strategies and added derived rules to each of these strategies:

- Exact matches
- Specific to general
- General to specific
- Many to one
- Cyclic mapping strategies
- No matches
- Specific and broad class mapping

6.3.4 CARMENx

The Universitätsbibliothek Regensburg leads the mapping efforts between different classification systems in the CARMEN project. Some members of the Renardus DDC mapping core group visited Regensburg on the 18 January 2001 and were shown the following mapping language. It has been used for some time already by staff from Regensburg (Regensburger Verbundklassifikation RVK) and the IZ Sozialwissenschaften in Bonn (Thesaurus structure of IZ Bonn Social sciences thesaurus) and is implemented in the mapping tool CARMENx (cf. 8.3).

- Identity ("=")
- Broader term ("<")
- Narrower term (">")
- Similarity ("<")

In the case of similarity one of three levels of relevance has to be chosen: high, medium or low. These "relevance" levels indicate degrees of overlap.

Broader and narrower term are interpreted as true supersets resp. subsets, similarity means overlap similarity means other forms of overlap.

This mapping language was developed in Regensburg by expressing relationships between classes as mathematical set relations (to be illustrated in Venn diagrams). This is seen as important in a retrieval system related to different expected levels of precision and especially recall.

As can be seen when comparing, the Renardus solution adopted so far follows CARMENx to a large degree. Renardus does not use "relevance" levels but introduces two well-defined degrees of overlap (major and minor).

Several of the relationships developed for thesaurus mapping are not applicable for classification mapping, others are just formulated differently. The proposal of Iyer translates well to most of the relationships used by CARMENx and Renardus. Since Renardus (and CARMENx) use only one relationship between a given pair of classes the single to multiple equivalencies relevant to classification mapping from Iyer, Doerr and the ISO standard are split into several relationship pairs.

An extended experience with real classification mapping and the usage of it in browsing and retrieval services will show if these selected relationships are correct and sufficient for classification mapping.

7 Main issues to be dealt with in the mapping work, including some initial decisions

Already before we start the real mapping work a great number of important issues which have to be discussed and decided upon at one stage or another can be foreseen. Below is a somewhat structured list of the issues we are aware of at the moment, including a few decisions the DDC mapping core group already made.

We expect frequent additions and changes to this list throughout the projects lifetime.

Decisions made will be reflected in the living guideline document that directs the activities of the mapping staff.

7.1 Specifics of the usage of DDC in Renardus

1. Should the structure of the DDC be displayed in the Renardus browsing pages as it is in the official version of the classification system or does it need to be adapted to a) strict logic and hierarchy, b) European content, or c) the real content of the Renardus subject gateways? Comments: The agreement with OCLC does not allow us to change the display of the structure of the DDC. There is no evidence for the need of logical revisions yet (as there was with the adaptation of the ETH Zürich UDC in GERHARD). A need for changes could occur especially in history and geography related areas of the gateways content. Subject areas

heavily represented or missing in Renardus will not be moved in the DDC structure. This might be dealt with at the layout level (cf. 8.2 Usage). We watch the arguments and the efforts of the Norwegian national Subject Gateway at BIBSYS on the creation of a browsing structure on top of the DDC.
Decisions: We do not introduce structural changes at the moment. Mapping staff lists areas where they see an important need for change.

2. Is it possible to "glue" parts of specialised classifications into lower level DDC classes if the DDC is not fine-grained enough?
Comments: The agreement with OCLC does not allow us to change the display of the structure of the DDC. The issue of add-ons is not dealt with. Indirectly, a jump to a local subject gateway could sometimes reveal local further subdivisions of a mapped class which have not been mapped to the DDC.
Decisions: Mapping staff announces the potential need for such an operation to the core group.

3. Can captions of DDC classes be changed?
Comments: The agreement with OCLC does allow us to change captions in order to adapt them to European vocabulary, provided the content of the class described with the new vocabulary remains basically the same. Changes in the DDC captions needed to improve end-user browsing in an Internet service featuring digital documents like Renardus are not expressively allowed. Several Internet usages of the DDC have experimented with changed captions and so does OCLC itself.
Decisions: Mapping staff announces wishes for changed captions to the core group. This group keeps a common list of the proposals, decides which changes are introduced to Renardus in accordance with the agreement and discusses other cases with OCLC.

4. Some of the subject areas being in the focus of a gateway are located at a rather deep level in the DDC. How to solve the conflict between the often compact discipline structures in the specialised subject classifications and the shattering of the same discipline in the huge universal systems? (The field of Engineering is expressed in 800 classes in the specialised Ei classification system, but shattered in about 2300 categories in the DDC).

7.2 Usage and display of the mapping in Renardus

1. What will the outcomes of the mapping effort look like in the Renardus browsing pages?
Comments: The basic structure is the DDC hierarchy. Underneath many classes links will appear to offer jumping to related classes/directories in local gateways in case they are mapped to the DDC class in question. Different mapping relationships will be indicated with the links (cf. ch. 6). The general user interface for the browsing structure, how the structure is displayed, what kind of navigation is possible etc. will be developed together with Renardus WP2.
Decisions:

2. Should empty classes be displayed for the sake of the logical consistency and completeness?
Comments:
Decisions: Empty classes in the three top levels of the DDC will remain visible in the browsing structure. Further down we intend to declare them at the nearest higher level as empty and remove completely empty levels from the display.

3. Can the number of records in each class be displayed in the Renardus browsing pages?

4. Should changed captions/structure be indicated?

5. Should it be indicated in case a gateway further subdivides the deepest DDC class mapped to?

6. Can we display multilingual versions of DDC captions in the Renardus browsing pages?
Comments: At least the top three levels of the DDC could be available in several different languages.
Decisions: Discuss this issue again with OCLC.

7. What will be the treatment and effect of the DDC mapping in the Renardus advanced subject searching?

7.3 Mapping Methods

7.3.1 Depth of mapping at both sides: the DDC and the local systems

1. Should the classification systems used in the originating Subject Gateways be mapped in their full depth to the common DDC cross-browsing system?
Comments: In principle: yes, if there are enough resources in the local class to make it worthwhile and if the gateways can afford it. This will of course be a stepwise process over time. The Renardus service does not need to wait until everything is mapped, since a further subdivision always remains available from the local class/directory that is mapped already.
Decisions: The mapping for the top levels of the local gateways should be finished and published first. Then the mapping staff should work their way down the local hierarchy in a balanced way. The core group will assure a balanced situation for the Renardus pages as a whole.
Only local classes containing more than two resources should be mapped. This means that mappings might be added or removed from Renardus at a later stage, in case the number of resources increases or decreases.

2. How deep should the common DDC cross-browsing structure be?
Comments: In principle, as deep as necessary in order to map to an equivalent class (cf. ch. 6). The agreement with OCLC prescribes no certain limitation. In practice, however, the core group could pragmatically define lower limits if the DDC mapping level becomes too deep and isolated in the Renardus DDC structure to provide a useful browsing offer to the users.
Decisions:

3. Should class contents too specific to be mapped to the DDC at an equivalent level be mapped to a higher level class?
Comments: cf. ch. 6.
Decisions: No.

4. Is it a suitable alternative to map to the three-level-DDC alone or to do both a three-level and a "as-deep-as-possible" solution?
Comments: cf. Ewa Nieszkowska's discussion paper
Decisions: No.

5. At what depth should gateways map that class their resources directly using the full DDC?
Comments: In principle above rules apply. We need to look at the classification praxis of BUBL, BIOME *et al*. in more detail in order to advice on a suitable mapping method. The individual resources are sometimes classed in a very specific and "individualizing" way, with built numbers and similar things that the Renardus browsing service will not use. There is however probably not a large problem as long as the gateways use the DDC structure to organise their browsing pages. This is not always the case, though. If not, it is preferable that the local browsing structure is mapped to Renardus and not the individual resources in the gateway.
Decisions:

7.3.2 Local classes which contain both generalities and specialities

1. What should be the mapping principle if a local class consists of both general resources and many specific resources in a subject area?

Comments: If not too many subcategories in DDC are involved with several special related records in the local class, mappings from these subcategories to the more general local class might be necessary. Otherwise the local gateway should be advised to further subdivide the general class before the mapping is used.
Decisions:

2. Should a more specific subclass in DDC be mapped to a more general class in the local system?
Comments: Normally not if a higher class in the same branch of DDC is mapped to this class. An exception is when some of the subclasses are mapped to other parts of the same local system or to other gateways.
Decisions:

3. Should a higher level DDC class be mapped to both the equivalent local class and all its subclasses?
Comments: No.
Decisions:

7.3.3 Exclusion of non-topical classes (auxiliary tables)

1. Should non-topical classes be excluded from the mapping?
Comments: The main Renardus browsing is limited to the ten main topical classes and its subdivisions. Where the main topics are subdivided by time, geography and similar from the main tables there will be a browsing access to these facets.
Decisions: Yes. Only classes directly mentioned in the official DDC version should be used for mapping. When necessary, we allow geographical and historic subdivision tables to be used in mapping of classes. We, however, decide at a later stage if and how to display it.

2. Is it allowed to use the method of "building numbers" in DDC classification in order to map a local class in a better way?
Comments:
Decisions: No. We do not build numbers. The only notations allowed in the mapping are the ones who already exist in the DDC version we work with.

3. Should DDC "interdisciplinary" notations be used?
Comments: Cf. Ewa Nieszkowska's discussion paper
Decisions:

4. Should document types be expressed in the DDC classification?
Comments: Document types should not be mapped to the DDC. Geological journals are to be mapped to the Geology class alone. In Renardus advanced search document type metadata can be searched independently from the subject element.
Decisions:

How to treat language subdivisions?
Comments:
Decisions: We do not allow mapping of language subdivisions in Renardus (e.g. an author writing in German language)

7.3.4 Number of allowed mappings?

1. Is there an upper limit of how many times a class can be mapped to related classes?
Comments:
Decisions: In order to reduce interdependency in the browsing system we recommend not to assign more than four mappings from a given DDC class to local classes at one gateway. Local classes which often are far broader than DDC classes at a reasonable browsing level might need to exceed the number of four mappings.

7.3.5 Mapping relationships

1. Should the mapping work express relationships between the classes mapped to each other?
Comments: cf. ch. 6 for reasons, definitions and the chosen mapping language.
Decisions: Yes, we assign a mapping relationship between each pair of classes.

2. Are there any implications of the fact that Renardus does the mapping only in one direction, from DDC to the local classes?

7.4 Recommendations to partners regarding their local classification systems and principles

1. Do we foresee changes in the gateways local classification praxis?
Comments: The Renardus project has obviously no formal right to invoke any changes at the level of the local gateways. As with metadata praxis and in other areas we hope that the close co-operation in Renardus, the focus on interoperability and the cross-browsing service will lead to improvements at the local level as well. Such changes might greatly improve the consistency, accuracy and usability of the mapping effort and of Renardus as a whole.
Decisions:

2. What are potential areas of improvements?
Comments:

- introduce IDs and notations to home-grown systems
- remove non-topical stuff from the main browsing structures
- correct logical errors and errors related to classification rules
- correct inconsistencies in the structures and classification praxis
- change the level of granularity of the local classification

If the local gateway wants to do some changes the best time to do it would be before or during the first mapping effort.
Decisions:

3. When is it advisable to use an established and maintained classification systems rather than home-grown solutions?

4. Should an established system be adapted and in that case how?

7.5 Mapping information: encoding and tools

1. What syntax should be used to encode the mapping information?
Comments: cf. ch. 6 and 8.4
Decisions:

2. Is a RDF encoding of the classification systems and the mapping information foreseen during the Renardus project?
Comments: It will at least be investigated.
Decisions:

3. Is there a tool available supporting the mapping work?
Comments: Most probably yes, an adaptation of CARMENx. Cf. ch. 8.3
Decisions:

7.6 Quality control and coordination of the mapping work

1. What influence has the subject overlap between the subject gateways on the mapping practice?
Comments: The core group has started to investigate the potential subject overlap between Renardus gateways (cf. ch. 4). Gateways with important overlaps should co-ordinate their work and watch the outcome. Hopefully the mapping tool will assist in this effort. The core group will watch overlapping areas for consistency.
Decisions:

2. What additional quality measures will be applied?
Comments: The core group will:

- check the consistency and correctness of mappings with the rules and regulations agreed upon
- do initial evaluations based on samples
- watch mapping solutions where a certain subject is a core topic in one gateway but a fringe area ("garbage collector") in others

Decisions:

3. Is there any further evaluation taking place?
Comments: Every subject gateway should of course evaluate the outcome of the mapping with its constituency. To evaluate the comprehensive cross-browsing service will be an important activity of Renardus WP5.
Decisions:

7.7 Update of the mapping

1. What are possible reasons for an update of the mapping information?
Comments: Besides changes during the main mapping effort the following reasons can invoke updates:

- every quarter of a year a new upgraded DDC file is delivered to Renardus
- the gateway changes and extends its browsing structures and classification policy
- revisions by the mapping staff in order to correct errors and inconsistencies

Our technical solution and the tool allow interactive and immediate changes in the mapping information by the responsible staff. The frequency of the normalisation at the gateway side and of the updates of the central browsing structure is very high and changes will be reflected almost immediately.
Decisions:

2. Is a versioning of the mapping necessary?
Comments: Probably not. The mapping should always be up-to-date and display only the latest version. During the Renardus project we aim to upgrade to the newest DDC quarterly version. Hopefully, OCLC will be able to indicate the changes to make our work easier.
Decisions:

3. Who will maintain mappings in the future?
Comments: The agreement with OCLC on the usage of the DDC ends with the Renardus project. In order to use the mapping outside of Renardus a bilateral agreement with OCLC is necessary. Hopefully, a future Renardus service will get a new agreement with OCLC and the

gateway partners at the time will continue to maintain the mappings. In case established classification systems are directly used the best solution in the long run would be to convince the owners and OCLC to take over responsibility for a mapping. This would be a sustainable solution for subject gateways. Renardus should develop a policy in this question.
Decisions:

7.8 Other issues

1. How to administer the updating and adaptation of the guideline document throughout the project's lifetime?

2. How to find the best trade-off between consistency, accuracy and usability in the Renardus cross-browsing service?

3. How can we keep our practical solutions suitable for automatic classification and mapping?

4. To what a degree can mapping between thesaurus descriptors or subject headings and the DDC be used as a replacement for real classification mapping?

5. Can we formulate further research and development issues emanating from this practically oriented project?

8 Technical implementation

This chapter provides a short description of the technical background and solutions behind the Renardus cross-browsing pages and the DDC mapping.

The file containing the enhanced DDC delivered by OCLC is used to construct the structure and framework of the browsing pages. Mapping information provided by the Renardus partners is then added to these pages and the individual classes to provide jumping addresses to related subject directories at the local subject gateways.

The mapping information needs to be delivered in a certain format that can be produced by a mapping tool adapted from CARMENx.
During the project alternative approaches to the encoding of vocabulary systems and mapping relationships will be investigated.

The DDC mapping information is used as part of the data in the Renardus advanced search Subject element as well.

8.1 Creation of the browsing pages with the enhanced DDC

The enhanced DDC (Ed. 21) is delivered by OCLC in 10 files: four reference files and 6 XML encoded data files. Among the reference files is a preliminary XML DTD, tag/attribute information and the top-level hierarchy. The master data file in XML is 50 MB large and is, in addition, presented as five sub-files (tables, schedules, manual notes, and summaries). It contains 25 500 main schedule entries (notations) and 35 700 different records.

Separately we received a file that contains the hierarchical relations between the classes in the main schedule.

Using these two files an initial complete hierarchical set of web pages is generated allowing a user to navigate through the DDC hierarchy. According to the agreement with OCLC only the DDC captions are displayed, not the notations. Layout and user interface solutions need still to be optimised, though.

When the mapping work is finished many of the completely empty branches in the lower part of this hierarchy will be removed from the display.

Since the mapping information is in a known place separate for each gateway we can harvest it and use it centrally while still keeping each of the gateways mappings separate (necessary if we are going to indicate in the browsing pages which gateway the user will end up in when following a link).

The mapping information can regularly be imported (harvested) to the pilot broker where a static datastructure is built.

Alternatives for producing the browsing pages:

1: Statically built pages - updated regularly

2: Dynamically built pages - generated on the fly when requested.

a. A static Perl datastructure containing the entire DDC hierarchy that is read each time a browsing page is accessed. It is used to extract the parts of DDC needed to build the requested page. Mapping tables are similarly stored and used. Drawback is that the datastructures becomes rather large (several MBs) and it takes time to read them into the program that generates the browsing page dynamically.

b. Alternatively the datastructures can reside in a database (mySQL) and be queried by the program that generates the browsing page dynamically.

2b is the most dynamic solution allowing updates to show up instantaneously in the user interface (provided we can use the CARMENx database tables for this purpose). On the other hand this complicates the Pilot as well as making it dependent on a database either locally or through the network.

8.2 Encoding: Syntax of the mapping information

The mapping information needs to be provided by the partners in a common defined format, as follows:

LN Local classification notation
LC Local classification caption
RL Relationship code
DN DDC notation
DC DDC caption
UR Local browsing URL for the class
NT Notes
:RE: Record end

Example:

LN: 400
LC: Civil Engineering
RL: =
DN: 624
DC: Civil engineering
UR: http://eels.lub.lu.se/ei/40.html
NT: discuss this with Martin
:RE:
LN and NT can be empty, :RE: is always empty. The fields should start at the beginning of a new line.

The relationship codes are defined in ch. 6.

The first experiences with the mapping and its usage in Renardus could very well lead to the possible abandonment of one or two relationships in the practical application.

It will be decided during the mapping workshop if it is preferable for logical reasons to start the format sequence with the Dewey related fields followed by the local ones, since the mapping logically points from DDC to the local classifications.

A file containing all mapping information from one subject gateway can be stored at a local Web server. The subject gateway (SBIG-ID) incl. the name of the classification should be mentioned once on top of the mapping file.

In case we manage to adapt the CARMENx mapping tool to the Renardus mapping effort this format will be produced by the tool from the entries made into it.

8.3 The mapping tool CARMENx

CARMENx is a tool developed at the University Library Regensburg in Germany for the CARMEN project [CARMEN]. It is tailor-made for the process of creating concordances/mappings between classification systems. The tool requires the free database software mySQL, an Apache web server, JavaScript and php scripts at the server side. Other database software could be used as long as there is a php interface available for them.

The software is Web-based and can be used with 4th generation Web-browsers. It allows co-operative work carried out by several users. Classification systems and concordance information can be kept separately at different servers.

Mapping relationships can be defined asymmetrically as well as bi-directionally.

The user interface consists of three main windows: one for displaying and navigating the origin classification, the other the target classification and the third one receives and displays the mapping information including the relationship and notes. The classifications can be displayed in a short hierarchy format (hierarchy information must be contained in the SQL database) and in a rich format with all notes etc.

Mapping relationships are introduced as links in both classification windows. One classification can be sequentially worked through during the mapping process.

There is a simple search function for the classifications. Via a search string in the URL a direct jump to a certain class can be accomplished.

Two members of the Renardus DDC mapping core group visited Regensburg in January 2001 for a demonstration and discussions. The tool was found useful for Renardus provided it could handle the rather complicated and large DDC file and that real distributed usability where all three SQL databases reside at different sites was possible and efficient.
In the meantime we managed to squeeze the DDC into a CARMENx SQL database and a first test with distributed usability was successfully.

The mapping tool incl. the database or databases with the Renardus mapping information will reside at SUB Göttingen, the DDC database will for legal reasons be placed at the site of the central Renardus system. Local classification systems could be stored at Göttingen or locally.

Further adaptation to Renardus will be necessary before we can use the tool: e.g. the user interface needs to be translated to English, classification system structures occurring at Renardus gateways need to be modelled in SQL databases, the mapping information and relationships used in Renardus need to be implemented etc. Renardus will only use mono-directional mapping from the DDC to the local classifications.

Creation and adaptation of php scripts and change of SQL database structures is needed.

Open questions are at the moment if:
a) an adaptation for the case that different gateways use the same classification locally is possible
b) all mappings from a certain DDC class can be displayed in the DDC window in spite of the fact that active mapping only is done to one local classification
c) the classification which is run through sequentially can be switched.

When the tool is in use, mapping information needs to be configured in the Renardus format (see above) in the SQL db used for its storage. Alternatively, a conversion step is needed. Perl scripts can be used to export data from the SQL database containing the mapping information to the central Renardus system.

There are several alternatives for exploring the total mapping information:

1: The present model assumes that each gateway somehow produces a mapping file in a certain format at a known location.

2: Each gateway does a mapping in the CARMENx tool. The mapping information is regularly exported in a certain format to a known location for each gateway. This requires that the subject gateway name (SBIG-ID) is added to each mapping by the tool or that two gateways are not allowed to use exactly the same local version of a classification. Otherwise the generation of browsing pages with links will be rather complicated and slow.

3. A combination of 1 and 2.

8.4 Usage of the mapping information in the browsing pages

The mapping information is taken from the SQL databases and used in the Renardus system to add mapping links to the browsing pages. The outcome on the Renardus browsing pages (notations will be omitted) could contain the following information:

Examples:
(691) Building materials
(691.96) Bituminous materials
Broader: EELS (411) `Bituminous materials`
Major overlap with: EELS (411.1) `Asphalt`

(627) Hydraulic engineering
(627.8) Dams and reservoirs
Narrower:
EELS (411.1) `Dams`
EELS (411.2) `Reservoirs`

(627) Hydraulic engineering
Minor overlap with: EELS (440) `Water and Waterworks Engineering`

8.5 Advanced subject searching with DDC

The DDC mapping information is used as part of the data in the Renardus advanced search Subject element as well after proper normalisation carried out by the gateways. Using the mapping information each gateway needs to add the DDC class to every single record in the gateway that belongs to the local class mapped to this DDC class.

The subject element combines both all local subject information (uncontrolled keywords, controlled keywords from thesauri and subject headings systems and classification captions and notations) with the captions of the mapped DDC classes.

In the pilot implementation a free text search on "words" is the only possible retrieval alternative. Phrase and word list searching will be added later. A search in the subject element opens normally a scan window that allows the user to select among subject entries hit by the search.

We will work on to make it possible to see the vocabulary schemas in the scan lists.

9 Related work

9.1 Projects

Apart from Renardus, there are a number of other current projects that are investigating the production of mappings between subject classification schemes, subject headings and thesauri. Two of them (i.e., CARMEN AP12, MACS) have developed from libraries' long experience of developing bibliographic databases; another from the social science data archive world (i.e., LIMBER). In short, the problem that they are attempting to solve is user focussed. Different libraries have traditionally implemented a range of different subject-indexing standards (subject schemes) for providing access to resources listed in their catalogues. These might include subject headings like MeSH or LCSH or classification codes like DDC or LCC. This means that those users who use more than one library catalogue have to learn the intricacies of each system. In addition, where natural language terms are used - e.g. in subject heading lists or thesauri - these tend to be in the language of the host institution, thus adding a further barrier for some users. The same issues arise (as we have seen) when bibliographic databases that contain incompatible subject schemes are merged together or combined into a union catalogue.

9.1.1 CARMEN AP12

CARMEN (Content Analysis, Retrieval and MetaData: Effective Networking) is a German project funded by the Global Info programme - a programme supported by German scientific societies, scientific information centres, libraries, publishing houses and the German Federal Ministry for Education and Research (BMBF). It is comprised of a number of different work-packages. Work-package 12 (AP12) of CARMEN concerns "Konkordanz von Klassifikationen und Thesauri" (Cross concordances of classifications and thesauri) and is led by the Universitätsbibliothek Regensburg. Other partners include IZ Sozialwissenschaften, Die Deutsche Bibliothek (DDB), the Deutsches Institut für Internationale Pädagogische Forchung (DIPF), the Max-Planck-Institut für Bildungsforschung, and Leske und Budrich publishers.

The work-package is concerned with the problem that where libraries use different subject indexing terms (including classification schemes), searching across them is impossible. The project Web pages (Schupfner, 2000) point out that the user "has to work with different search terms and the respective search logic of the system, so that an efficient search is hardly possible." The project first aims to examine the methodologies that might underlie the development of mappings between different classification schemes or thesauri (these are here called cross concordances). It will then develop procedures for representing these cross concordances and a prototype system.

Subject schemes that might be considered for the production of mappings by CARMEN AP12 would mainly be based on the areas of physics, mathematics, and the social sciences. The project documentation suggests that the main candidate schemes would appear to be the DDC, the Regensburger Verbundklassifikation (RVK), the Mathematics Subject Classification (MSC) developed by the American Mathematical Society and the Physics and Astronomy

Classification Scheme (PACS) developed by the American Institute of Physics. Thesauri and subject headings might include the Schlagwortnormdatei (SWD), the thesaurus of the IZ Sozialwissenschaften and the key word terms used by the DIPF (CARMEN project, 2000).

9.1.2 The LIMBER project

Project LIMBER (Language Independent Metadata Browsing of European Resources) is a project concerned with the development of multilingual access to the contents of social science archives. The project is partially funded by the European Union under its Information Society Technologies (IST) programme. Partners include the Central Laboratory of the Research Councils (CLRC) Rutherford Appleton Laboratory, the UK Data Archive, the Norwegian Social Science Data Services (NSD), Intrasoft and a number of other European social science data archives.

The LIMBER project will adapt the Humanities and Social Science Electronic Thesaurus (HASSET) - initially developed by the Data Archive - by extending it with 'equivalent terms' in non-English languages; initially in French, German and Spanish (Matthews and Wilson, 2000). A multilingual search tool will use these equivalent terms to search for relevant data across a number of data sources. Matthews and Wilson (2000) note that a major problem is that the 'equivalent term' in different languages may not have an exact match and that "this is particularly a problem in social science (as opposed to the physical sciences) where the exact meaning of terms is culturally dependent." They further note that this problem may get worse if the thesauri are arranged in hierarchies.

The LIMBER project itself builds on a completed Telematics Applications Programme project called NESSTAR (Networked Social Science Tools and Resources). This project was concerned with developing software tools that would facilitate the publishing and retrieval of statistical and other social science data on the Web. (Musgrave and Ryssevik, 2000). The NESSTAR system utilised an international standard for social science metadata, an XML-based 'codebook' developed by the Data Documentation Initiative (DDI). One of LIMBER's aims is to develop a multilingual thesaurus that can operate either on its own or as a plug-in to the NESSTAR system.

9.1.3 The MACS project

MACS (Multilingual Access to Subjects) is a project of the Conference of European National Librarians (CENL), financed by the Swiss National Library, the Bibliothèque nationale de France (BnF), Die Deutsche Bibliothek (DDB) and the British Library. The project is primarily concerned with the problem of multilingual access to bibliographic databases with subject headings in different languages. Clavel-Merrin (1999) has noted that the majority of subject searches in bibliographic databases "are only possible using the subject entries in the language of the country." As an example, she notes that in the Bibliothèque nationale de France, "each document, independently of the language in which it has been written, is indexed using a French-language subject heading language." Landry (2000, p. 2) describes the problem using the Swiss National Library as an example: "a researcher from abroad must not only master the German language but also the intricacies of the SWD/RSWK subject heading language in order to access material by subject."

Following a pilot study based on the subject heading languages used in the four partner institutions (LCSH, RAMEAU and SWD/RSWK), the libraries funded the development of a prototype MACS system that would establish links between the headings in each language. A consortium made up of Index Data (Denmark) and Tilburg University (Netherlands) was contracted to develop this prototype system. In order to test the prototype, c. 15,000 bibliographic records from each database and 1,000 headings were loaded into the system. These headings were in the fields of sport and theatre, together with an additional set of links derived from the 500 most-used RAMEAU headings (Landry, 2000. p. 3). Links between the different subject heading languages are dealt with by the Link Management Interface.

9.2 Literature

The linking of indexing vocabularies to each other is not a new idea. In the 1980s, some new subject thesauri were themselves based upon faceted library classification schemes. For example, the Bliss Bibliographic Classification (BC) was used to help create the DHSS-DATA and ECOT thesauri (Aitchison, 1986). Most current work on mapping relates to the work of abstracting and indexing services or library catalogues. Their main purpose is either to help with end-user searching - e.g. by providing a 'natural-language' path into the subject vocabularies used by the databases themselves - or to facilitate consistent information retrieval in an heterogeneous environment.

9.2.1 Abstracting and indexing services

Mapping between multiple subject schemes is a feature of some developments undertaken by abstracting and indexing services. Good examples are the Unified Medical Language System developed by the National Library of Medicine (NLM) and an application of MeSH in the BIOSIS databases (Hodge, 2000, pp. 18-19). Another is the OmniFile service provided by H.W. Wilson. This provides access to six of H.W. Wilson's databases by merging all the indexing terms into one file. The product publicity says, "diligent editing reconciles subject headings in the various specialties for uniformity throughout, so your search won't miss a single relevant citation."

The Unified Medical Language System (UMLS)

The Unified Medical Language System (UMLS) has been developed by the NLM in order to link various medical vocabularies into a single 'Metathesaurus.' The Metathesaurus itself is a human-created database of information on concepts that appear in over 60 controlled vocabularies and subject classifications. The *2001 UMLS Documentation* on the Web (NLM, 2001) notes that the Metathesaurus "includes about 800,000 concepts and 1.9 million concept names in different source vocabularies." It adds that it "preserves the meanings, attributes, hierarchical connections, and other relationships between terms present in its source vocabularies, while adding certain basic information about each of its concepts and establishing synonymy and new relationships between concepts and terms from different source vocabularies."

Because the UMLS contains information about the relationships between concepts, there has been some interest in demonstrating how the system could be used to help provide automatic translations from one medical terminology to another (e.g., Cimino, *et al*., 1993). Also, other researchers have found the UMLS useful for the mapping of non-MeSH concepts to MeSH (Bodenreider, *et al*., 1998) and for automatic classification of condition terms into broad disease categories (Bodenreider, 2000).

The use of MeSH by BIOSIS

Another example of an abstracting and indexing service providing mappings relates to co-operation between BIOSIS and the NLM. BIOSIS contributes records to the NLM's TOXLINE database to which appropriate MeSH terms are automatically added. Hodge (2000, p. 18) describes how:

> *This is based on a mapping of the natural language terms that occur in the toxicology literature and BIOSIS' normalised natural language keyword indexing with the MeSH terminology. In the new BIOSIS relational indexing structure, BIOSIS builds and maintains authority files that connect natural language disease names to the MeSH-controlled disease terms. When the BIOSIS indexer assigns the free text keyword for the disease name, the appropriate MeSH term is also added to the record as an alternate access point ... The assignment is based on the development over time of a mapping between the terminology used by BIOSIS and the MeSH-controlled terms.*

Hodge (2000, p. 18) further notes that the existence of MeSH terms in the BIOSIS databases would facilitate the development of services that would be able to cross-search both BIOSIS and medical databases like MEDLINE. It also might help support the use of the BIOSIS databases by biomedical researchers and medical practitioners already familiar with using MeSH terms.

9.2.2 Library catalogues

Chan (2000) notes that "because the networked environment is open and multifarious, multiple tools for resource description and subject access are often used side-by-side." She notes that the use of multiple subject vocabularies within the same system is not uncommon. In order to provide consistent subject access to a bibliographic database, there is sometimes a need for mappings between different schemes. One published example is a mapping between Library of Congress Subject Headings (LCSH) and Medical Subject Headings (MeSH) produced by a project based at Northwestern University (Olson and Strawn, 1997). Chan (2000) gives examples of more established mappings, including the linking of some LCC codes in the LCSH lists and the inclusion of abridged DDC numbers in the Sears List of Subject Headings. Other examples include links made between DDC and LCSH terms by OCLC (Vizine-Goetz, 1998b).

ALCTS subcommittees

In the United States, subcommittees set up by the Subject Analysis Committee of the Cataloging and Classification Section of the Association for Library Collections & Technical Services (ALCTS/CCS/SAC) have undertaken focussed studies. In 1996, the SAC set up a Subcommittee on Subject Relationships/Reference Structures. This subcommittee was set-up to investigate the various relationships that exist between subjects and to make recommendations on how these relationships should be recorded in standard formats and how they should be presented to users. The subcommittee found a large number of semantic relationships and recommended a subset as potentially being useful for information retrieval (ALCTS/CCS/SAC/Subcommittee on Subject Relationships/Reference Structures, 1997).

Another subcommittee, the Subcommittee on Metadata and Subject Analysis, was charged in 1997 to "identify and study the major issues surrounding the use of metadata in the subject analysis and classification section of digital resources." A report from the subcommittee was published in July 1999. This made certain recommendations for the use of classification schemes in Web metadata. The report recommended that methods for the harmonisation of subject terms from different controlled vocabularies should be developed and refined. It pointed to the development of services like the UMLS Metathesaurus, and recommended the "investigation of the feasibility of developing a general metathesaurus or expanding the medical metathesaurus to include indexing terms covering all subject areas" (ALCTS/CCS/SAC/Subcommittee on Metadata and Subject Analysis, 1999, 3.2.4).

University of California, Berkeley - Entry Vocabulary Modules

Various strands of research carried out at the University of California, Berkeley has looked at the automatic mapping of entry vocabularies to unfamiliar metadata vocabularies. Buckland, *et al*. (1999) note that catalogues and databases often use unfamiliar indexing terms and categorisation schemes. Effective searching in a heterogeneous context is often compromised by the existence of these unfamiliar terms and codes.

> *When an index or categorization scheme is encountered, how is one to know what word or value has been assigned to the topic that one is interested in?*

The answer is to create a search module that can provide guidance from a familiar vocabulary to an unfamiliar scheme. To do this, the Berkeley researchers developed something called an Entry Vocabulary Module based on association dictionaries (Norgard, 1998). The examples that were developed by the Berkeley team were based on mapping natural language queries to

unfamiliar indexing terms, although Buckland, *et al*. (1999) note that the same approach could be used for mapping between different vocabularies. The Berkeley researchers have developed English language indexes to BIOSIS Concept Codes, the INSPEC Thesaurus and the US Patent and Trademark Office Patent Classification (Buckland, et al., 1999). They also developed a multilingual index to LCC.

The point of the research at Berkeley is not to depend on human expertise to establish the links (or mappings) between the entry and target vocabularies. Unlike human-moderated systems like the UMLS, Buckland, *et al*. (1999) intend to produce low-cost automatically-generated links. The Berkeley team use a technique known as "classification-clustering" to create ranked lists of relevant terms in a target vocabulary (Larson, 1991).

9.2.3 Some German developments

Classification for scientific libraries

An expert group was founded in Germany in January 1995 to investigate the usage and further developments of classification systems in German scientific libraries (Deutsches Bibliotheksinstitut, 1998). The following questions were in the foreground:

- Are existing classification systems suitable for future usage in German scientific libraries?
- In which way should existing classification systems be developed further for OPACs and the usage in library networks like the GBV (the Gemeinsamer Bibliotheksverbund of the German states of Bremen, Hamburg, Mecklenburg-Vorpommern, Niedersachsen, Sachsen-Anhalt, Schleswig-Holstein and Thüringen)?
- Is it possible to develop a concordance between several classification systems?
- The expert group took only universal classification systems into consideration. They also checked only existing classification systems, their goal was not to develop a new system.
- Several classification systems were investigated regarding:
    - historical aspects, genesis
    - application and coverage of application/geographical extension,
    - general structure: general systematic structure, emphasis on certain aspects like the, so called, "White Anglo-Saxon Protestant" orientation, actuality, maintenance etc.,
    - structure (main classes, tables etc.), depth, structure of notations
    - conclusions

Three popular world-wide classification systems were studied by the expert group: the Dewey Decimal Classification (DDC), the Universal Decimal Classification (UDC), and the Library of Congress Classification (LCC). Several popular German classification systems, e.g., the Basic Classification (Basisklassifikation, BK), the Subject Authority File (Schlagwort-Normdatei, SWD) - a classification system used by the DDB - and the Regensburger Verbundklassifikation (RVK), were also studied.

The expert group recommends in their final report the use and dissemination of the Dewey Decimal Classification and the development of cross-concordances to DDC. The main precondition of this will be a German translation of the DDC. The DDB has decided to start a German translation at the end of 2001.

Ingo Nöther's model of an international concordance-classification

From 1993 until 1996 Ingo Nöther from the Ehemaliges Deutsches Bibliotheksinstitut (EDBI; the former DBI) was involved in a project to build a model of an international concordance-classification (Nöther, 1998). The goal of the project was to develop a model of a concordance-classification in order to adjust the subject indexing to automation and globalisation. The study focused on the subjects "general" (including library sciences, bibliography, museology), "arts", and "music". The following classification systems were used to build up an international concordance-classification:

- Dewey Decimal Classification (DDC) – Edition 20
- Universale Dezimalklassifikation (UDK) – International Middle Edition
- Bibliothekarisch.Bibliographische Klassifikation (BBK) – Russian Edition
- Regensburger Verbundklassifikation (RVK)
- Aufstellungssystematik der Gesamthochschulbibliotheken des Landes Nordrhein-Westfalen (GHB-Systematik) (the shelf classification of the University Libraries of North Rhine-Westphalia)
- Basic Classification (BK)

The goal was not to develop a concordance on the basis of a so-called "register-concordance" or "document-concordance" but with the aid of some kind of "hyper-concordance". In this way relations between different classes of several classification systems and the hyper-classification will be produced. With the help of the hyper-classification a concordance between the classes (Systemstellenkonkordanz) could be developed. The "Systemstellenkonkordanz" and the underlying hyper-classification together form the concordance-classification.

The development of a concordance between one classification system and another system or other systems does not take place in a direct way. The concordance will be expressed indirectly via the hyper-classification (the master). The advantages of using a master and to map from one classification system to a master are logical: it is not necessary to develop several concordances between different classification systems, only one concordance from each classification system to the hyper-classification must be developed. Via one single hyper-classification all concordances between different classification systems are covered. This master or concordance-classification must fulfil several conditions: It must be possible to develop concordances from all classes of the several classification systems to the classes of the concordance-classification. Furthermore the concordance-classification should be a universal, flexible, analytically structured, and strongly hierarchical one.

Nöther developed a tool for demonstration purposes. The concordance-classification is based on DDC (edition 20).

9.2.4 Other subject mappings

DDC-UDC mapping in SOSIG

With more direct relevance to Renardus, at least one mapping between classification schemes has been produced by subject gateways. The Institute for Learning and Research Technology (ILRT) at the University of Bristol undertook a limited mapping between the Universal Decimal Classification (UDC) and the Dewey Decimal Classification (DDC) in order to facilitate the use of SOSIG metadata records by the Biz/ed service. In 1999, the browse-structure of the SOSIG Internet catalogue contained 16 top-level subject areas with over 230 sub-sections (Hiom, 2000, p. 54). This browse structure was based on UDC codes that were

added to each metadata record. When the Biz/ed service was being set up, it was felt useful to share all records with the SOSIG database. The Biz/ed service, however, also wanted to use a browse-structure based on DDC, which was perceived by subject experts as being more up-to-date in the business studies area than UDC. As part of the DESIRE project, ILRT constructed a mapping file that provided a translation between the relevant UDC and DDC numbers. This was used to convert the Biz/ed DDC codes into UDC when templates (i.e., metadata records) were periodically copied from Biz/ed to the SOSIG database. The original DDC codes were not thrown away, but additional subject-descriptor fields would be automatically created to hold the newly created SOSIG UDC numbers (Hiom, 1998).

It is worth noting that subject gateways do not always use the codes. A mapping between DDC and a visual browse interface will be used in the Norwegian BIBSYS portal (Husby, 2000).

The Aquarelle project's Thesaurus Management System

Aquarelle was a project funded by the European Commission under its Telematics Applications Programme from 1996-1998 (Michard, 1998). The project was concerned with developing an Information Network on Cultural Heritage. This is a distributed information system that attempts to provide uniform access to the varied collections of data held by museums, art galleries and other cultural heritage-based organisations within Europe.

One of the Aquarelle project's outcomes was the Semantic Index System-Thesaurus Management System (SIS-TMS). This is an application of the Semantic Index System (SIS) produced by Aquarelle partner ICS-FORTH. SIS-TMS is a tool to support the development of multilingual thesauri and a "terminology server for cataloguers and for distributed access to heterogeneous electronic collections" (Michard, 1998). Doerr and Fundulaki (1998, p. 15) say that SIS-TMS allows one "to maintain multiple, multilingual thesauri and their interrelations in one logical database." The system has a graphic user interface that permits the understanding and control of complex interlinked terminology structures (Doerr, 1998).

10 References

Aitchison, J., 1996, A classification as a source for a thesaurus: the Bibliographic Classification of H. E. Bliss as a source of thesaurus terms and structure. *Journal of Documentation*, 42 (3), 160-181.

ALCTS/CCS/SAC/Subcommittee on Metadata and Subject Analysis, 1999, *Subject data in the metadata record: recommendations and rationale*.
`http://www.ala.org/alcts/organization/ccs/sac/metarept2.html`

ALCTS/CCS/SAC/Subcommittee on Metadata and Subject Analysis:
`http://www.ala.org/alcts/organization/ccs/sac/metasub.html`

ALCTS/CCS/SAC/Subcommittee to Promote Subject Relationships/Reference Structures, 1997, *Final Report to the ALCTS/CCS Subject Analysis Committee*.
`http://www.ala.org/alcts/organization/ccs/sac/rpt97rev.html`

ALCTS/CCS/SAC/Subcommittee to Promote Subject Relationships/Reference Structures:
`http://www.ala.org/alcts/organization/ccs/sac/srrs.html`

Arbeitsgruppe Klassifikatorische Erschliessung, 2000, *Einführung und Nutzung der Dewey Decimal Classification (DDC) im deutschen Sprachraum*. Frankfurt am Main: Die Deutsche Bibliothek.

Batty, D., 1998, WWW - wealth, weariness or waste: controlled vocabulary and thesauri in support of online information access. *D-Lib Magazine*, November.
`http://www.dlib.org/dlib/november98/11batty.html`

Bodenreider, O., 2000, Using UMLS semantics for classification purposes. *Proceedings of the AMIA 2000 Annual Symposium*, 86-90.
http://www.medicine.ucsd.edu/f2000/D200831.htm

Bodenreider, O., Nelson, S.J., Hole, W.T., Chang, H.F., 1998, Beyond synonymy: exploiting the UMLS semantics in mapping vocabularies. *Proceedings of the AMIA 1998 Annual Symposium*, 815-819. http://www.medicine.ucsd.edu/f98/D004771.htm

Brickley, D., 1998, *Rule-based metadata crosswalks using RDF: a case study using classification scheme mapping*, v. 1.0. Bristol: ILRT, 16 April.
http://www.ilrt.bris.ac.uk/discovery/rdf-dev/purls/papers/classmap/

Buckland, M., Chen, A., Chen, H.-M., Kim, Y., Lam, B., Larson, R., Norgard, B., Purat, J., Gey, F., 1999, Mapping entry vocabularies to unfamiliar metadata vocabularies. *D-Lib Magazine*, 5 (1), January.
http://www.dlib.org/dlib/january99/buckland/01buckland.html

Buckland, M.K., Butler, M.H., Norgard, B.A. Plaunt, C., 1992, OASIS: a front-end for prototyping catalog enhancements. *Library Hi Tech*, 10 (4), 7-22.

CARMEN AP12: http://www.bibliothek.uni-regensburg.de/projects/carmen12/

CARMEN project, 2000, *Crosskonkordanzen von Klassifikationen und Thesauri*. CARMEN AP12. http://www.mathematik.uni-osnabrueck.de/projects/carmen/AP12.html

CARMEN project: http://www.mathematik.uni-osnabrueck.de/projects/carmen/carmen.html

Chan, L.M., 2000, *Exploiting LCSH, LCC, and DDC to retrieve networked resources: issues and challenges*. Bicentennial Conference on Bibliographic Control for the New Millennium, Library of Congress, Washington, D.C., 15-17 November 2000.
http://lcweb.loc.gov/catdir/bibcontrol/chan_paper.html

Cimino J.J., Johnson, S.B., Peng, P., Aguirre, A., 1993, From ICD9-CM to MeSH using the UMLS: a how-to guide. In: Safran, C., ed., *Proceedings of the Seventeenth Annual Symposium on Computer Applications in Medical Care*. New York: McGraw-Hill, 730-734.

Clavel-Merrin, G., 1999, The need for co-operation in creating and maintaining multilingual subject authority files. 65th IFLA Council and General Conference, Bangkok, Thailand, 20-28 August 1999. http://www.ifla.org/IV/ifla65/papers/080-155e.htm

Cross, P., Day, M., Koch, T., Peereboom, M., Zettergren, A. -S., 1999, Subject classification, browsing and searching. In: Belcher, M., Knight, V., Place, E., (eds.), *DESIRE Information Gateways Handbook*. DESIRE II deliverable D3.4.
http://www.desire.org/handbook/2-5.html

Deutsches Bibliotheksinstitut, 1998, *Klassifikationen für wissenschaftliche Bibliotheken: Analysen, Empfehlungen, Modelle*. (dbi-materialien, 175). Berlin: Deutsches Bibliotheksinstitut.

Dewey Decimal Classification: http://www.oclc.org/fp/

Dewey-Dezimalklassifikation (DDC): Infoseite der DDB:
http://www.ddb.de/professionell/ddc_info.htm

Doerr, M., 1998, *Effective terminology support for distributed digital collections*. Sixth DELOS Workshop, Tomar, Portugal, 17-19 June 1998. Le Chesnay: ERCIM, pp. 53-58. PDF version available at: http://www.ercim.org/publication/ws-proceedings/DELOS6/index.html

Doerr, M., 2000a, *Thesaurus mapping*. MODELS 11 workshop, Bath, UK, 11 January. PowerPoint presentation available at: http://www.ukoln.ac.uk/dlis/models/models11/doerr.ppt

Doerr, M., 2000b, *Thesaurus mapping*. ECDL 2000, European Conference on Digital Libraries, Lisbon, Portugal, 20 September. http://www.bn.pt/org/agenda/ecdl2000/NKOS2.htm

Doerr, M., 2001, Semantic problems of thesaurus mapping. *Journal of Digital Information*, (forthcoming). http://jodi.ecs.soton.ac.uk/

Doerr, M., Fundulaki, I., 1998, The Aquarelle Terminology Service, *ERCIM News*, 33, 14-15.

Geißelmann, F., 1998, *Classifications in Germany. State of the discussion at the scientific libraries*. Paper presented at a workshop held at DDB.

Facharbeitsgruppe Sacherschließung des GBV, 2000, *Basisklassification*, 3., erweiterte Ausgabe. Göttingen: Gemeinsamer Bibliotheksverbund. http://www.gbv.de/sacher/bk-info.htm

Gey, F., Chen, H.-M, Norgard, B., Buckland, M., Kim, Y., Chen, A., Lam, B., Purat, J., Larson, R., 1999, *Advanced search technologies for unfamiliar metadata*. IEEE Meta-Data '99: the third IEEE Meta-Data Conference, Bethesda, Md., 6-7 April 1999. http://computer.org/proceedings/meta/1999/papers/36/fgey.html

Heiner-Freiling, M., 1998, DDB und DDC - Die Deutsche Bibliothek und die Dewey-Dezimalklassification. *Bibliotheksdienst*, 32 (12). http://www.dbi-berlin.de/dbi_pub/bd_art/98_12_10.htm

Hiom, D., 1998, *Mapping classification schemes*. DESIRE project. Bristol: Institute for Learning and Research Technology, February. http://www.sosig.ac.uk/desire/class/mapping.html

Hiom, D., 2000, SOSIG: an Internet hub for the social sciences, business and law. *Online Information Review*, 24 (1), 54-58.

Hodge, G., 2000, *Systems of knowledge organisation for digital libraries: beyond traditional authority files*. Washington, D.C.: Council on Library and Information Resources, Digital Library Federation, April. http://www.clir.org/pubs/abstract/pub91abst.html

Hood, M.W., Ebermann, C., 1990, Reconciling the CAB Thesaurus and AGROVOC. In: *Quarterly Bulletin IAALD*, 35 (4), 181-185.

Husby, O., 2000, *Emneportaler: en gjennomgang av kjennetegn og oppbygging med sideblikk til BIBSYS emneportal*. Emneportaler til Internett: Mangfold og samordning, Georg Sverdrup-seminarene 2000, Oslo University Library, 20 November 2000. MS PowerPoint slides available from: http://www.jbi.hio.no/bibin/kurs/portal/

Iyer, H., Giguere, M, 1995, Towards designing an expert system to map mathematics classificatory structures. *Knowledge Organization*, 22 (3-4), 141-147.

Kim, Y., Norgard, B., 1998, *Adding natural language processing techniques to the entry vocabulary module building process.* Technical report.
http://www.sims.berkeley.edu/research/metadata/nlptech.html

Koch, T., 2000, Quality-controlled subject gateways: definitions, typologies, empirical overview. *Online Information Review*, 24 (1), 24-34.
http://www.lub.lu.se/~traugott/OIR-SBIG.txt

Koch, T., Day, M., 1997, *The role of classification schemes in Internet resource description and discovery*. DESIRE deliverable D3.2 (3).
http://www.ukoln.ac.uk/metadata/desire/classification/

Koch, T., Vizine-Goetz, D., 1998, Automatic classification and content navigation support for Web services: DESIRE II cooperates with OCLC. *Annual Review of OCLC Research, 1998*. Dublin, Ohio: OCLC.
http://www.oclc.org/oclc/research/publications/review98/koch_vizine-goetz/automatic.htm

Landry, P., 2000, *The MACS Project: Multilingual Access to Subjects (LCSH, RAMEAU, SWD)*. 66th IFLA Council and General Conference, Jerusalem, Israel, 13-18 August 2000.
http://www.ifla.org/IV/ifla66/papers/165-181e.pdf

Larson, R.R. 1991, Classification clustering, probabilistic information retrieval and the online catalog. *Library Quarterly*, 61 (2), 133-173.

LIMBER project: http://venus.cis.rl.ac.uk/limber/

Lorenz, B., 1994, Notizen zum Stand klassifikatorischer Arbeit. Ein Diskussionsbeitrag. (Antwort auf den Beitrag von Ingo Nöther). *Bibliotheksdienst*, 28 (6), 870-878

MACS project: http://infolab.kub.nl/prj/macs/

Mandel, C.A., 1987, *Multiple thesauri in online library bibliographic systems*. Washington, DC: Library of Congress, Cataloging Distribution Service.

MARC 21 Concise format for classification data. (2000):
http://lcweb.loc.gov/marc/classification/eccdhome.html

Markey, K., 1989, Subject searching strategies for online catalogues through the Dewey Decimal Classification. In: Hildreth, C.R., ed., *The online catalogue: developments and directions*. London: Library Association, 61-83.

Matthews, B., Miller, K., 2000, *Modelling the LIMBER thesaurus in RDF*. (Manuscript).

Matthews, B., Wilson, M., 2000, *Multilingual metadata to access social science data*. Data Management Workshop, CLRC Daresbury Laboratory, Cheshire, UK, 23-25 February.
http://venus.cis.rl.ac.uk/limber/External/paper1.htm

McKiernan, G., 1996-2001, *Beyond bookmarks: schemes for organizing the Web*. Ames, Iowa: Iowa State University. http://www.iastate.edu/~CYBERSTACKS/CTW.htm

Meo-Evoli, L., Negrini, G., Farnesi, T, 1998, ICC and ICS: Comparison and Relations between Two Systems Based on Different Principles. In: el-Hadi, W.M., Maniez, J., Pollitt, A.S., eds., *Structure and relations in knowledge organization: proceedings of the 5th International ISKO Conference, Lille, 25-29 August 1998*. Würzburg: Ergon Verlag, pp. 229-237.

Michard, A., ed., 1998, *Final report: IE-2005 Aquarelle: sharing cultural heritage through multimedia telematics*. Le Chesnay: INRIA. http://aqua.inria.fr/Aquarelle/Public/EN/final-report.html

Miksa, F.L., 1998, *The DDC, the universe of knowledge, and the post-modern library*. Albany, N.Y.: Forest Press.

Mitchell, J., 1995, DDC 21 and beyond: the Dewey Decimal Classification prepares for the future. *Cataloging and Classification Quarterly*, 21 (2), 37-48.

Mitchell, J.S., Vizine-Goetz, D., 2000, *A research agenda for classification*. Dublin, Ohio: OCLC. http://www.oclc.org/dewey/research/research_agenda.htm

Mitchell. J., *et al*., ed., 1996, *Dewey decimal classification and relative index: devised by Melvil Dewey*, 21st ed. 4 vv. Albany, N.Y.: Forest Press.

Musgrave, S., Ryssevik, J., 2000, *Beyond NESSTAR: faster access to data*. IASSIST 2000 - Data in the Digital Library: Charting the Future for Social, Spatial and Government Data, Northwestern University, Evanston, Ill., USA, 7-10 June. Draft PDF version available at: http://www.nesstar.org/papers/beyondnesstar.pdf

National Library of Medicine, 2000, *Fact Sheet: UMLS Metathesaurus*. http://www.nlm.nih.gov/pubs/factsheets/umlsmeta.html

National Library of Medicine, 2001, *2001 UMLS knowledge sources*, 12th ed. Bethesda, Md.: NLM, January. http://www.nlm.nih.gov/research/umls/UMLSDOC.HTML

NESSTAR project: http://www.nesstar.org/

Nöther, I. 1994, Modell einer Konkordanz-Klassifikation für Systematische Kataloge. *Bibliotheksdienst*, 28 (1) 15-33 und 28 (2), 175-187.

Nöther, I., 1998, Zurück zur Klassifikation!: Modell einer internationalen Konkordanz-Klassifikation. In: *Klassifikationen für wissenschaftliche Bibliotheken: Analysen, Empfehlungen, Modelle*. (dbi-materialien, 175). Berlin: Deutsches Bibliotheksinstitut, 103-325.

Norgard, B., 1998, *Entry vocabulary modules and agents*. Technical report. Berkeley, Ca.: School of Information Management and Systems, University of California, Berkeley. http://www.sims.berkeley.edu/research/metadata/agents.html

Olsen, H., 1998, Mapping beyond Dewey's boundaries: constructing classificatory space for marginalized knowledge domains. *Library Trends*, 47 (2), 233-254

Olson, T., Strawn. G., 1997, Mapping the LCSH and MeSH systems. *Information Technology and Libraries*, 16 (1), 5-19.

OmniFile: http://www.hwwilson.com/databases/omnifile.cfm

Plaunt, C., Norgard, B.A. 1998, An association based method for automatic indexing with a controlled vocabulary. *Journal of the American Society for Information Science*, 49, 888-902. http://bliss.berkeley.edu/papers/assoc/assoc.html

Schupfner, G., 2000, *WP12: Cross concordances of classifications and thesauri*. Regensburg: CARMEN project. http://www.bibliothek.uni-regensburg.de/projects/carmen12/

SIS-TMS: Semantic Index System - Thesaurus Management System: http://www.ics.forth.gr/proj/isst/Systems/sis-tms.html

Soergel, D.,1999, The rise of ontologies or the reinvention of classification. *Journal of the American Society for Information Science*. 50 (12), 1119-1120.

Studwell, W.E., 2000, USE, the Universal Subject Environment: a new subject access approach in the time of the Internet. *Journal of Internet Cataloging*. 2 (3/4), 197-209.

Stumpf, G., Internet -Informationen zur Sacherschließung, 5. Klassifikatorische Sacherschließung. http://www.bibliothek.uni-augsburg.de/allg/swk/sacher5.html

Svenonius, E., 2000, *The intellectual foundation of information organization*. Cambridge, Mass.: MIT Press.

Sweeney, R., 1983, The development of the Dewey Decimal Classification. *Journal of Documentation*, 39 (3), 192-205.

Vizine-Goetz, D., 1996a, Using library classification schemes for Internet resources (position paper). *Proceedings of the OCLC Internet Cataloging Colloquium, San Antonio, Texas, January 19, 1996*. Dublin, Ohio: OCLC. http://www.oclc.org/oclc/man/colloq/v-g.htm

Vizine-Goetz, D., 1996b, Online classification: implications for classifying and document-like object retrieval. Electronic version of a paper published in: Green, R., ed., *Knowledge organization and change: proceedings of the Fourth International ISKO Conference, 15-18 July 1996, Washington, DC, USA*. Frankfurt/Main: Indeks Verlag. http://orc.rsch.oclc.org:6109/dvgisko.htm

Vizine-Goetz, D., 1997a, OCLC investigates using classification tools to organize Internet data. *OCLC Newsletter*, no. 226, March/April. http://www.oclc.org/oclc/new/n226/research.htm#investigates

Vizine-Goetz, D., 1997b, From book classification to knowledge organization: improving Internet resource description and discovery. *Bulletin of the American Society for Information Science*, 24 (1), 24-27. http://www.asis.org/Bulletin/Oct-97/vizine.htm

Vizine-Goetz, D., 1998a, Dewey as an Internet subject guide. Electronic version of paper published in: el-Hadi, W.M., Maniez, J., Pollitt, A.S., eds., *Structure and relations in knowledge organization: proceedings of the 5th International ISKO Conference, Lille, 25-29 August 1998*. Würzburg: Ergon Verlag, pp. 191-197. http://www.oclc.org/~vizine/isko/vizine-goetz_isko5.htm

Vizine-Goetz, D., 1998b, Subject headings for everyone: popular Library of Congress Subject Headings with Dewey numbers. *OCLC Newsletter*, no.233, May/June. http://www.oclc.org/oclc/new/n233/rsch_subj_headings_everyone.htm

Vizine-Goetz, D., Godby, D., 1997, Library classification schemes and access to electronic collections: enhancement of the Dewey Decimal Classification with supplemental vocabulary. In: Solomon, P., (ed.), *Advances in classification research, Vol. 7: proceedings of the 7th ASIS SIG/CR classification research workshop, 20 October 1996, Baltimore, Maryland*. Medford, N. J.: Information Today, pp. 127-35. http://www.oclc.org:5047/~vizine/sig_cr/sigcr_done_dvg.htm