DESIRE: Project Deliverable

Project Number:	RE 1004 (RE)	
Project Title:	DESIRE - Development of a European Service for Information on Research and Education	
Deliverable Number:	D3.2	
Deliverable Title:	Specification for resource description methods Part 3. The role of classification schemes in Internet resource description and discovery.	
Version Number	1.0	
Deliverable Type:	PU	
Deliverable Kind:	RE	
Principal Authors:	Name	Traugott Koch
	Address	Lund University Library, Development Department NetLab, P.O. Box 3, S-221 00 Lund, Sweden.
	E-Mail	traugott.koch@ub2.lu.se
	Telephone	+46 46 2229233
	Fax	+46 46 2223682
	Name	Michael Day
	Address	UKOLN, University of Bath, Bath BA2 7AY, UK.
	E-Mail	M.Day@ukoln.ac.uk
	Telephone	+44 1225 323923
	Fax	+44 1225 826838
Other Authors:	Anna Brümmer, Debra Hiom, Marianne Peereboom, Alan Poulter, Emma Worsfold.	
Deliverable URL(s):	http://www.ub.lu.se/desire/radar/reports/D3.2.3/class_v10.html http://www.ub.lu.se/desire/radar/reports/D3.2.3/class_v10.rtf	
Abstract:	Classification schemes have a role in aiding information retrieval in a network environment, especially for providing browsing structures for subject-based information gateways on the Internet. Advantages of using classification schemes include improved subject browsing facilities, potential multi-lingual access and improved interoperability with other services. Classification schemes vary in scope and methodology, but can be divided into universal, national general , subject specific and home-grown schemes. What type of scheme is used, however, will depend upon the size and scope of the service being designed. A study is made of classification schemes currently used in Internet search and discovery services, particular reference being given to the following schemes: Dewey Decimal Classification (DDC); Universal Decimal Classification (UDC); Library of Congress Classification (LCC); Nederlandse Basisclassificatie (BC); Sveriges Allmäma Biblioteksförening (SAB); Iconclass; National Library of Medicine (NLM); Engineering Information (Ei); Mathematics Subject Classification (MSC) and the ACM Computing Classification System (CCS). Projects which attempt to apply classification in automated services are also described including the Nordic WAIS/WWW Project, Project GERHARD and Project Scorpion.	
Keywords:	Classification schemes; Resource discovery; Resource description; Internet;	

The role of classification schemes in Internet resource description and discovery

Contributors:

Anna Brümmer, NetLab, Lund University Library, Sweden

Michael Day, UKOLN, University of Bath, UK

Debra Hiom, Social Science Information Gateway, University of Bristol, UK

Traugott Koch, NetLab, Lund University Library, Sweden

Marianne Peereboom, Koninklijke Bibliotheek, The Hague, Netherlands

Alan Poulter, Department of Information and Library Studies, Loughborough University, UK

Emma Worsfold, Social Science Information Gateway, University of Bristol, UK

PART II: Executive Summary

Some Internet services concerned with giving access to other Internet sites use classification schemes for organising a browsing structure giving access to their selected resources. This is especially true of Internet subject services which often use a browsable classified structure in addition to a searchable index.

The use of a classification scheme gives some advantages to an Internet service to the extent that it helps with browsing, enables the broadening and narrowing of searches, gives context to search terms being used, allows (under certain conditions) multi-lingual access to collections of material and the partitioning and manipulation of a large database. If an exisiting classification scheme is chosen, it will have a good chance of not becoming obsolete and will possibly be well-known to users.

Classification schemes can be defined by several categories, but can be broadly divided into:

- Universal schemes - examples include the Dewey Decimal Classification (DDC), the Universal Decimal Classification (UDC) and the Library of Congress Classification (LCC);
- National general schemes - universal in subject coverage but usually designed for use in a single country. Examples include the Nederlandse Basisclassificatie (BC) and the Sveriges Allmäma Biblioteksförening (SAB);
- Subject specific schemes - designed for use by a particular subject community. Examples include Iconclass for art resources, the National Library of Medicine (NLM) scheme for medicine and Engineering Information (Ei) for engineering subjects;
- Home-grown schemes - schemes devised for use in a particular service. An example from the Internet is the 'ontology' developed for the *Yahoo!* search service.

All of these classification types are used to some extent on the Internet. Universal schemes like DDC and UDC are used by many Internet services and are readily available in machine-readable form. Subject services, however, appear more likely to use a subject specific scheme.

The type of classification scheme chosen for use in an Internet service should depend upon the scope of service which is planned. A subject service, where possible, could use a well-known, international, subject specific scheme. Another service, which either has a more general brief or is in a subject areas where there is no agreed 'standard' classification system in use, could use or adapt a unversal scheme.

For the widest interoperability, more than one classification scheme could be used or conversion programs designed. Alternatively, a universal scheme could be used to 'glue' different subject services together while the actual services themselves would be classified in a different, relevant, subject-specific scheme.

Classification is a time-consuming and expensive process, so research has been carried out into the automatic classification of Internet resources. Various projects have investigated how subject terms collected from a search of a database can be converted into classification notation. Two projects, the Nordic WAIS/WWW project (Lund) and Project GERHARD (Oldenburg) used UDC for the conversion, while OCLC's Project Scorpion is looking at DDC. Other projects are looking at neural-networks and at automatic conversions between classification schemes.

Automatic classification processes are also important if large robot-generated services want to add a browsing structure for their documents.

Contents:

PART III:
The role of classification schemes in Internet resource description and discovery

1. Introduction and overview

1.1. Background

This report investigates the use of classification schemes to aid retrieval in a network environment, specifically with regard to the Internet. The library community, over many years, had appeared to favour subject indexing systems (the use of a controlled vocabulary to assign indexing terms to documents) over the use of traditional classification schemes (grouping documents into a hierarchical structure of subject categories). During the first period of the development of networked information services, many specialists, especially those from the computing community, also questioned the value of library subject description systems in principle, pointing to the accomplishments of full-text indexing software.

The increasing use of the Internet and the World Wide Web (WWW) for the storage and retrieval of vast amounts of information has, however, changed this perception. Two distinct ways of finding resources on the Internet emerged (Dodd 1996, p. 276). One approach consisted of the development of robot based search engines which could be used for powerful keyword searches of the contents of the WWW. These are extremely useful tools, although they have a tendency to return large amounts of irrelevant information. The other approach started with producing 'hotlists' which would encourage users to browse the WWW. The production of hierarchical browsing tools sometimes led to the adoption of library classification schemes to provide the subject hierarchy. At least one general discovery service, *Yahoo!* <URL:http://www.yahoo.com/>, devised their own 'home-grown' classification scheme (or ontology) to give structured hierarchical access to the resources which they had indexed. Quality-controlled subject services, which gave access only to selected Internet resources, also understood that a browsing structure based on subject classification would be a desirable compliment to a search engine type service. Most subject services of this type, and almost all of the Electronic Libraries (eLib) Programme access to network resources services and the proposed DESIRE test-bed services currently use a classification scheme which can be browsed. A list of Internet sites that use library classification systems or subject headings can be found in *Beyond bookmarks* (McKiernan 1996) <URL:http://www.iastate.edu/~CYBERSTACKS/CTW.htm>.

This report will describe the advantages of resource classification for subject-based information gateways in the Internet and will analyse the advantages and disadvantages of different types of classification systems and will then review some important individual schemes.

1.2. Advantages and disadvantages of classification

The use of classification schemes offers one solution to providing improved access to WWW resources. Web sites have been created to act as a guide to other Web sites selected according to some pre-specified criteria, e.g. they are judged to be good quality resources or relevant to a particular subject-area. Some of these sites typically consist of an alphabetical list of subjects, and selected Web resources are listed below each one.

Examples include *Argus Clearinghouse* <URL:http://www.clearinghouse.net/> and the *WWW Virtual Library* <URL:http://www.w3.org/pub/DataSources/bySubject/Overview2.html>. In this context, it can be understood why classification schemes have begun to be used to give added-value subject access to Web sites. A site that organises knowledge with a classification scheme demonstrates several advantages over sites which do not (cf. Svenonius 1983):

- Browsing: classified subject lists are easily able to be browsed in an online environment. Browsing is particularly helpful for inexperienced users or for users not familiar with a subject and its structure and terminology. In addition, the structure of the classification scheme can be displayed in different ways as a navigation aid. 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: classification schemes are hierarchical and therefore can be used to broaden (i.e. for improved recall) or narrow a search when required. Questions can be limited to individual parts of a collection (filtering) and the number of false hits be 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 (words which have the same form and spelling but a different meaning) can be partly overcome.
- Potential to permit multilingual access to a collection: since classification systems often use notations independent from a specific language, indices in different languages can offer multilingual access to the same resources without any further changes to the collection. A searcher could enter search terms in a given language and those terms would then relate to the relevant parts of the classification system (as a switching language) and be used to retrieve resources in any given language on the subject.
- The partitioning and manipulation of a database: large classified lists can be divided logically into smaller parts if required.
- The use of an agreed classification scheme could enable improved browsing and subject searching across databases.
- An established classification system is not usually in danger of obsolescence. The larger schemes are now undergo 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.
- They 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 an Internet service which uses them will therefore have an advantage over one that uses its own classification or none.
- Many classification schemes are available in machine-readable form.

Classification schemes, however, can be sometimes subject to criticism:

- 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.
- The illogical subdivision of classes: some popular schemes do not always subdivide classes in a logical manner (Buchanan 1979, pp. 32-34; Rowley 1987, pp. 188-189). This can make them difficult to use for browsing purposes.
- Assimilating new areas of interest: classification schemes, since they are usually updated through formal processes by organised bodies, often reveal difficulty in reacting to new areas of study.

1.3. The typology of classification schemes

1.3.1. Types of classification systems

There are several different types of classification systems around, varying in scope, methodology and other characteristics. Detailed descriptions cannot be given here, but it might be useful to know these different types, when trying to understand the terminology of this report and when decisions about which scheme to use is required.

Classification systems - by facet:

- by subject coverage: general or subject specific
- by language: multilingual or individual language
- by geography: global or national
- by creating/supporting body: representative of a long-term committed body or an home-grown system developed by a couple of individuals

- by user environment: libraries with container publications or documentation services carrying small focused documents (e.g. abstract and index databases)
- by structure: enumerative or faceted
- by methodology: a priori construction according to a general structure of knowledge and scientific disciplines or using existing classified documents

(The categories are not dichotomic, a classification can fit into more than one category).

The facet structure above shows what types of classification scheme are theoretically possible. In reality, the most frequently used types of classification schemes are: a) universal; b) national general; c) subject specific schemes, most often international; d) home-grown systems; d) local adaptations of all types.

The term 'universal' schemes is used for schemes which aim to include all subjects, are global geographically and multilingual in scope. Part 2 of the report deals with some of the most well-known individual schemes as examples.

1.3.2. Universal classification schemes

The first practical universal classification schemes were developed in the late-nineteenth-century as a response to the problem of organising libraries in the context of rapidly growing knowledge and an increase in the numbers of printed books. Universal schemes aim to be both comprehensive and also to expand and contract to fit the state of knowledge at any time.

The most widely-used universal classification schemes are those which were developed for the use of libraries since the late-nineteenth-century, notably the Dewey Decimal Classification (DDC), the Universal Decimal Classification (UDC) and the classification scheme devised by the Library of Congress (LCC).

Use of a universal, multidisciplinary classification scheme in an Internet context results in the following advantages (in addition to the general advantages of using a classification scheme, see 1.2 above):

- They can cover all subject areas: The use of an agreed universal classification scheme as a global top-level structure could enable improved browsing and subject searching across services and collections from different subject areas. In theory, the use of an agreed universal scheme at many sites would allow for the widest interoperability. But it should be remembered that this is normally not the most important criteria when choosing a scheme for a certain service (cf. 4. Conclusions).
- They are widely supported: For the universal schemes, there is a global interest in support, development and survival of the scheme. DDC, UDC and LCC have been repeatedly revised since their first publication and are updated by responsible international bodies.
- They might be known to more users than other types of classifications: regular users of libraries will be familiar with at least part of one or more of these schemes. Use of an Internet service which uses them will therefore have an advantage over one that uses its own classification or none.
- They have an especially good potential to permit multilingual access to a collection: DDC was first published in English and UDC in French, but have both been widely translated. Full editions of UDC have been made available in English, German, Russian and Spanish, and abridged versions are available in other languages (Langridge 1973, p. 89; McIlwaine and Buxton 1995, pp. 7-8). DDC has been translated into 30 languages and is currently used in 135 countries (Thompson, Shafer & Vizine-Goetz 1997). This means that the tools already exist for multilingual access to Internet sites organised with these schemes.
- The major universal classification schemes are now all available in machine-readable form (see parts 2.1 - 2.3)

Universal classification schemes, however, are subject to several criticisms:

- False ontology: there is a general concern that universal schemes impose a false order upon knowledge. For example it was believed in the early 1970s that DDC still reflected its origins in a small North American university library (Foskett 1973, p. 39). The structure of enumerative schemes (most universal schemes are basically enumerative) is often perceived as subjective, and critics find many examples of inconsistency and illogicality. For this reason, library classification theory had begun to move away from enumerative schemes in the mid-twentieth-century. Examples of the alternative 'faceted' or 'analytico-synthetic' classification schemes are Ranganathan's Colon

classification (Ranganathan 1965) and Bliss's Bibliographic classification (although both are hardly ever used), although later editions of DDC and UDC are faceted to a limited extent.

- Bad at assimilating new areas of interest: universal classification schemes often have a special difficulty in reacting quickly to new areas of study because they are updated with the time consuming participation of broad international multidisciplinary bodies. Researchers on the University of Illinois Digital Library Initiative project comment that most digital repositories contain "concepts and vocabularies too new or dynamic for controlled-vocabulary-based human indexing" (Schatz, B. *et al*. 1996, p. 33). Similarly, all classification systems are poor at handling new concepts and vocabularies, but universal classification schemes tend to have more disadvantages in this area when compared with subject-specific schemes.

1.3.3. National general schemes

Most of the advantages and disadvantages of universal classification schemes apply also to national general schemes (cf. 2.4. National general schemes), but they have additional characteristics that make them perhaps not the best choice for an Internet service that claims to be relevant for a wider user group than one limited to certain national boundaries.

Some of those characteristics are discussed here, relating to use of the scheme in the Internet environment:

- Although national general schemes offer coverage of all subject areas, they are in general not well known outside of their place of origin. For an international audience, a universal scheme would probably serve better.
- Support for a national scheme will be broad in the country itself, and a national institution has the responsibility for development. Support for the scheme outside of this national user group is limited. (e.g. use of the Nederlandse Basisclassificatie by German libraries which use the Pica system).
- Within the country the national scheme may be better known than universal schemes, e.g. the BC is used by Pica libraries in the Netherlands (mostly academic libraries), and SAB is used by almost all the public libraries in Sweden.

When the choice was made in the Koninklijke Bibliotheek to use the Nederlandse Basisclassifatie for an Internet subject service (the *Nederlandse Basisclassificatie Web*), this was done mainly because the subject specialists already used the scheme for classification of printed works. If *NBW* outgrows its national boundaries, for instance in the DESIRE context, or by the participation of non-Dutch institutions, the conversion to another scheme will deserve serious consideration, to make wider interoperability possible.

- Multilingual capability is not a primary concern for national schemes, apart from countries with multiple languages.
- National schemes are likely to have a geographic bias, e.g. the classification of languages in the BC is not only Eurocentric, but biased towards the Dutch context: Frisian - as a language spoken by a minority in Holland - has a separate class, while Asiatic languages have only three: Japanese, Chinese and 'other' Asiatic languages. This bias could be a serious drawback in an international context.

1.3.4. Subject specific schemes

Most special subject specific schemes have been devised with a particular user-group in mind. Typically they have been developed for use with indexing and abstracting services, special collections or important journals and bibliographies in a scientific discipline. They do have the potential to provide a structure and terminology much closer to the discipline and can be more up-to-date, compared to universal schemes.

Examples of specific schemes are Engineering Information (Ei) for engineering, the National Library of Medicine (NLM) Classification for medicine and the British Catalogue of Music Classification. In subject areas like medicine, agricultural science and engineering, where there are international and widely recognised schemes available, subject services normally will prefer these or use them in combination with an universal scheme.

Subject specific schemes do have some drawbacks:

- It makes co-operation between subject services from different subject areas more difficult. Elaborate conversion programs will be needed in order to exchange resources or to point to them in another service.
- If they have a very small user-base it can be very difficult for the numerous users from other subject areas to learn the structure of the scheme.
- Collections of subject-specific resources are likely to include some fringe topics which will not be adequately covered within the specialist scheme itself (Langridge 1991, p. 16)..

It is therefore advisable that only well-established subject specific classification schemes should be used to describe Internet resources.

1.3.5. Home-grown schemes

Some Web sites have tried to organise knowledge on the Internet by devising their own classification scheme. *Yahoo!*, created in 1994, lists Web sites using their own universal classification scheme or 'ontology', which contains 14 main categories. Each Web site collected for *Yahoo!* is listed under one of 20,000 categories or sub-categories (Steinberg 1996), the scheme being developed over time by the 20 people doing the classification work.

A study by Vizine-Goetz (1996a) showed that out of *Yahoo!'s* 50 most popular categories, all but four mapped perfectly to explicit DDC or LCC numbers or ranges. The results "... indicate that DDC and LCC have sufficiently wide topic coverage for classifying Internet resources". The structure of *Yahoo!* would require encoding to take advantage of the relationships between classes which is handled by notations in traditional library schemes, an important prerequisite for automatic routines and improved navigation.

Home-grown schemes do have some theoretical advantages over library universal classification schemes:

- Home-grown schemes are relatively flexible and easy to change. For example, in 1995 *Yahoo!* was adding categories and making other changes to the ontology every day (Steinberg 1996).
- Home-grown schemes can very quickly absorb new areas of interest. Universal and enumerative schemes cannot just add new classification numbers when they are required, attention has to be given to keeping the numeric arrangement logical and easy to understand. This process can be very drawn-out.

On the other hand, home-grown schemes have a number of disadvantages:

- They amplify the problems of classification subjectivity and can lead to a lack of consistency. Steinberg (1996) notes that *Yahoo!*'s more or less consistent point of view "comes from having the same 20 people classifying every site, and by having those people crammed together in the same building where they are constantly engaged in a discussion of what belongs where". Other people using the same scheme or ontology might come to very different solutions.
- They are unlikely to be as well-known to users as universal classification schemes.
- If the scheme is self devised, it might need frequent revision with little chance of co-operation. The economic cost of this will fall entirely on the originator of the scheme.

2. Current use of classification schemes in existing search services

2.1 The Dewey Decimal System (DDC)

The Dewey Decimal Classification System (DCC) was first produced by Melvil Dewey in 1876, originally being produced for a small North American college library. It is currently in its 21st edition (Mitchell 1995; Mitchell, *et al.* 1996) and is published by Forest Press. DDC is distributed in Machine-Readable Cataloguing (MARC) records produced by the Library of Congress (LC) and bibliographic utilities like OCLC and RLIN. DDC is also used in the national bibliographies of the UK, Canada, Australia, Italy and other countries (Comaromi, et al. 1990, p.6). Research carried out by OCLC in the 1980s established that DDC was a suitable tool for browsing, first for library catalogues and then for the Internet (e.g. Markey 1989; Vizine-Goetz 1996a).

2.1.1. Extent of usage in Internet services

McKiernan (1996) lists 14 sites that use, or claim to use, DDC for the organisation of resources. 6 of these sites were also available from the relevant part of the *Yahoo!* ontology: *Computers and Internet:Internet:World Wide Web:Searching the Web:Indices to Web Documents:Dewey Decimal Classification*
<URL:http://www.yahoo.com/Computers_and_Internet/Internet/World_Wide_Web/Searching_the_Web/Indices_to_Web_Documents/Dewey_Decimal_Classification/index.html>.

The *Yahoo!* page included a reference to the Champaign Public Library page and McKiernan (1996) included the Utah State Library but there appeared to be no Web pages organised by DDC at either of these library sites. These 2 sites were ignored. 13 sites remained:

- *Basalt Regional Library Homepage.* Basalt, Colo.: Basalt Regional Library. <URL:http://www.colosys.net/basaltlib/>
- *Blue Web'n Content Categories*. San Diego, Calif.: Pacific Bell. <URL:http://www.kn.pacbell.com/wired/bluewebn/categories.html>
- *Canadian Information by Subject*. Ottawa: National Library of Canada.<URL:http://www.nlc-bnc.ca/caninfo/esub.htm>
- *CyberDewey: a Catalogue for the World Wide Web*. David A Mundie (Pittsburgh, Penn.). <URL:http://ivory.lm.com/~mundie/DDHC/CyberDewey.html>
- *GNOSIS (Global Network Of Silicon Information Services)*. Patrick W. Clancey (Sun Microsystems). <URL:http://www.aimnet.com/~clancey/gnosis/>
- *The Internet Resource*. Edinburgh: Napier University Library. <URL:http://www.napier.ac.uk/depts/library/intres/ir000999.html>
- *Internet Resources in Dewey Decimal Order with DDC Subjects*. Independence, Mo.: Mid-Continent Public Library <URL:http://www.mcpl.lib.mo.us/dewey.htm>
- *Mr. Dui's Topic Finder*. Dublin, Ohio: OCLC Office of Research. <URL:http://www.oclc.org/oclc/fp/mrdui/mrdui.htm>
- *Net Sites by the Numbers* Tempe, Ariz.:Tempe Public Library. <URL:http://www.tempe.gov/library/bnindex.htm>
- *Sites by DDC*. Alastair G. Smith (Victoria University of Wellington, New Zealand). <URL:http://www.vuw.ac.nz/~agsmith/bkmark.htm#ddc>
- *WEBrary(tm)* (Online Ready Reference System). Morton Grove, Ill.: Morton Grove Public Library. <URL:http://www.nslsilus.org/mgkhome/orrs/webrary.html>
- *World Wide Web Reference Collection*. Denton, Tex.: University of North Texas, School of Library and Information Sciences <URL:http://www-lan.unt.edu/slis/student/projects/wwwrc/Index.htm>

- *WWLib Browse Interface* (Wolverhampton Web Library). Peter Burden (University of Wolverhampton, UK) <URL:http://www.scit.wlv.ac.uk/wwlib/browse.html>

An Internet search was also undertaken for the terms 'Dewey Decimal' or 'DDC'. Web pages were searched using the major robot-based search engines (*AltaVista, Excite, HotBot, InfoSeek/UltraSeek, Lycos, Webcrawler*), and searches were made of *USENET* postings (via *Dejanews*) and the archives of the *PACS-L* and *web4lib* mailing lists. The following 2 extra sites were found:

- *Expanding Universe*. Toronto, Ontario: Metropolitan Toronto Reference Library. <URL:http://www.mtrl.toronto.on.ca/centres/bsd/astronomy/index.html>
- *PICK: Quality Internet Resources in Library and Information Science* Aberystwyth: University of Wales Aberystwyth, Thomas Parry Library. <URL:http://www.aber.ac.uk/~tplwww/e/pick.html>

Expanding Universe specialised in astronomy resources, and *PICK* in resources for information and library studies.

Two further Internet sites that use DDC are *NetFirst* and *Biz/ed*:

- *NetFirst*. Dublin, Ohio: OCLC. <URL:http://www.oclc.org/oclc/netfirst/>

NetFirst has used DDC to organise a browsing structure since October 1996. DDC notations had been present in their links from the start of the service, but have only recently been made available for browsing (Oehler 1996).

- *Biz/ed*. Bristol: University of Bristol, Institute for Learning and Research Technology <URL:http://www.bized.ac.uk/>

Biz/ed is a subject service for business education, based at the University of Bristol and funded by a consortium of commercial and educational organisations, including Elib. *Biz/ed* offers an online catalogue of good quality Internet resources (using the ROADS software, and based on the *SOSIG* model - in fact, *SOSIG* and *Biz/ed* share the same offices). Since its inception in 1996, this catalogue has used an abridged version of DDC to classify resources, and to create browsable subject sections.

NetFirst can be considered a major search tool, and *Biz/ed* aims to develop into an important resource for those interested in business education. The other sites were on a smaller scale, although *PICK*, *Expanding Universe* and the *Canadian Information by Subject* sites were useful because of the specialised nature of the resources they gave access to. The access figures for those sites which had them did not indicate heavy use.

The sites can be categorised into those maintained by libraries and universities, and those maintained by dedicated individuals Another categorisation relates to the type of resources held, either general or subject specific. All the sites used DDC to order resources, with the exception of the *Blue Web'n*, a site specialising in Web resources for education. Its use of DDC was not as an organiser but as an adjunct to a list of subjects in alphabetical order. Each subject was given its 3-digit DDC number, merely as a reference.

Of the general-resource, library-based sites, *Basalt Regional Library*, *Mid-Continent Public Library* and *Morton Grove Public Library* (also known as *WEBrary*) used primarily 3-digit DDC numbers to group the resources they pointed to. Even at this general level of classification many numbers only led to single resources. The *WEBrary* was notable for including a short review alongside each resource. The *Internet Resource* at Napier University Library was the most threadbare in terms of the number of resources pointed to. Many DDC numbers had no resources attached.

The best of the general-resource, library-maintained sites was the National Library of Canada's *Canadian Information by Subject*. Here DDC numbers went beyond the initial decimal point and the quantity of resources meant that clumps of resources appeared under individual DDC numbers. Nevertheless the classification was not that deep.

Of the specific-resource, library/university-based sites, two of the sites made full use of the DDC, using number-building facilities appropriate for their focused domains: *Expanded Universe*, based at the Metropolitan Toronto Reference Library, *PICK* at the Thomas Parry Library, University of Wales Aberystwyth, These sites used the longest DDC numbers of all the sites surveyed. Biz/ed had used the business section of DDC to pick out a selection of numbers and classes that could be used to form the browsable sections on the site.

Of the sites maintained by individuals, three stand out as being excellent guides to general resources, *Global Network Of Silicon Information Services (GNOSIS)* run by Patrick W. Clancey, *CyberDewey: A Catalogue for the World Wide Web* run by David A. Mundie and *WWLib* by Peter Burden at the University of Wolverhampton. *CyberDewey* is perhaps the weakest of the three, with many resources being pages on *Yahoo!*, *BUBL* or the *World Wide Web Virtual Library* and mainly shorter DDC numbers being used. Both *GNOSIS* and *WWLib*, despite being general resource collections, contain some long and well-built DDC numbers, especially in areas like computing, presumably of interest to their creators. Burden, the creator of *WWLib*, is also an advocate of using automated techniques to classify the Web (Wallis and Burden 1995).

Alastair Smith's *Bookmarks on the Net* is what it says it is, his somewhat eclectic bookmarks organised using short DDC numbers. The *World Wide Web Reference Collection* purports to organise reference sources by DDC but is almost devoid of resources.

In conclusion, the two specialised collections, *Expanded Universe* and *PICK* are perhaps the most successful as DDC gives them the facility to organise tightly-focused resources. The *Biz/ed* service has pragmatically used a section of DDC to structure the browsable interface of the catalogue. The general collections have the feeling of overlong bookmarks, with DDC less in evidence as a powerful organiser.

2.1.2. Extent of usage in traditional and other online services

DDC is used by more libraries than any other classification scheme. It is currently used in 135 different countries and has been translated into 30 languages (Thompson, Shafer & Vizine-Goetz 1997). It is used by the Library of Congress Cataloguing Service in the bibliographic records it creates, alongside the Library of Congress Classification (LCC). DDC is found in many national bibliographies and in the records created by the major commercial bibliographic services. DDC classifications also appear (together with LCC and Library of Congress Subject Headings (LCSH))in the Cataloguing in Publication (CIP) data produced by the Library of Congress, the British Library and some other national libraries. OCLC is now the owner and publisher (via Forest Press) of DDC and the company maintain a Web page for Forest Press and DDC at: <URL:http://www.oclc.org/oclc/fp/fptxthm.htm>

2.1.3. Multilingual capability

The famous DDC 'decimal' notation, because it consists solely of digits and decimal points, is language independent. There is an attempt in DDC to attach meaning to groups of digits, albeit in a somewhat unwieldy manner.

Translations of DDC exist in French, Italian, Spanish, Turkish and other languages. For more information about current products see: <URL:http://www.oclc.org/oclc/fp/products/fpprod-t.htm>.

2.1.4. Strengths and weaknesses of the scheme

The first edition of DDC was published in 1876. As such it was a product of its time and was imbued with a Western, Christian, pre-technological age viewpoint. Subsequent revisions (the latest being Edition 21 (Mitchell *et al*. 1996)) have done much to try and alleviate this. For example, Edition 21 saw a major revision of the life sciences in 560-590 which saw a move away from the organism centred approach in earlier versions of DDC towards an orientation concentrating more upon processes (New and Trotter 1996). The scheme is updated more frequently than other universal schemes.

Engineering

The 620s have long been a problem in DDC, as the process of revision struggles to keep up with the tangled growth of engineering. To give one example, building, as a human act that involves design, is spread between 624, 690 and 720. From Dewey's time this trifurcation has been a problem.

Art

DDC is used successfully by several large art libraries, however it must be said that the entire 700 class has been faulted for its fragmentation and overlapping, but criticisms tend to focus on the final two divisions. Moves to revise the 780 schedule have not met with success as it is still an extremely difficult schedule to use while the 790s are still chaotic.

Social sciences

The 300s have seen much revision to try and iron out weaknesses, yet they still remain. For example, social groups are still classed separately from their culture. While the statistics of a subject are now classed at the subject with 021 appended, no number exists for the statistics of neo- or perinatal death or indeed, any mortality statistics with respect to a particular disease. The law schedule has seen major disagreement over whether jurisdiction or type of law should be classed first and, as a result, it allows 'options' in its interpretation.

Business and economics

Biz/ed compared the business and economics coverage of DDC with UDC and found in favour of DDC. This subject field has changed in emphasis in recent years (e.g. there is an increased emphasis on market economies), and it was found that the DDC reflected this change best, having been revised and updated more frequently that the UDC.

Using a universal scheme for a subject specific service can cause problems however, as the more detailed the classification, the more complicated it is to use the scheme.

2.1.5. Integration between classification scheme and other systems e.g. controlled subject headings

DDC numbers are linked to LCSH headings by most major bibliographic services to the extent that their bibliographic records contain LCSH headings together with DCC and LCC classification data. The USMARC record, for example, contains specific tags for several different classification schemes: DDC, LCC, UDC and NLM together with tags for subject headings including LCSH and MeSH.

DDC 21 classification numbers have also been mapped to LCSH, using statistical mappings between the two systems generated from the OCLC database.

Selected new LCSH headings are individually linked to DDC numbers and are made available via: <URL:http://www.oclc.org/oclc/press/961206.htm>, although this only includes a very small proportion of the complete LCSH listings.

2.1.6. Linking DDC to third party classification data

The USMARC format allows for links to be made between DDC and other classification systems, including LCC, UDC and NLM.

2.1.7. Digital availability

The DDC classification is maintained and edited in machine-readable format. The scheme is maintained using an Editorial Support System which was first used to create Edition 20 in 1990 (Comaromi *et al.*1989). Subsets of this database have been made available under licence for certain projects.

An online summary of the first 3 digits of DDC 21 numbers is available on the WWW at: <URL:http://www.oclc.org/oclc/fp/ddc/ddcsum21/ddc21sm1.htm>. The scheme has been made available to Internet users on an experimental basis with the intention of encouraging authors and other relevant people to use DDC to classify materials on the Internet.

The Software system Dewey for Windows is available on CD-ROM (for details see: <URL:http://www.oclc.org/oclc/fp/deweywin/dwytoc-t.htm>), however tools like this and the *Classification Plus* component of *Cataloger's Desktop* are only designed to give cataloguers' access to DDC and are not suitable for the application of DDC in an Internet environment.

2.1.8. Copyright

Copyright rests with Forest Press/OCLC. They can be contacted via the Forest Press page <URL:http://www.oclc.org/oclc/fp/welcome/fpwelc-t.htm>. Those using the classification would be able to apply the notation without restraint in, say, library catalogues and WWW pages, but use of the other information in the schedules would require permission from Forest Press.

2.1.9. Extensibility and development effort of the scheme

Edition 21 (Mitchell *et al* 1996) is the latest revision, produced under the direction of Forest Press. Although DDC has undergone constant and sometime radical revision from within it has remained structurally recognisable and constantly useful.

When devised, DDC was an enumerative scheme. Subsequent revisions have absorbed and made use of the significant challenges and contributions of classification theory, chiefly the structure and methodology of faceted classification and the use of facet analysis. As a result, subsidiary tables and 'divide like' devices that reflect and can express many aspects of complex topics have been expanded, even though DDC is not a faceted scheme.

DDC stands ideologically or theoretically between its two major counterparts: decidedly more flexible than the Library of Congress Classification and certainly simpler than UDC.

2.1.10. Other issues

DDC is well supported by an extensive explanatory and training literature. Its simple notation has accumulated considerable user familiarity over the long span of its existence.

Dewey editors are trying to code the nature of the relationships between class numbers and subject headings to see if it is possible to develop classifier and retrieval assistance tools (Vizine-Goetz, 1996b).

2.2. Universal Decimal Classification (UDC)

The Universal Decimal Classification (UDC) is an international scheme which endeavours to cover all areas of knowledge. Its origins lie in the Dewey Decimal Classification (DDC) which was adapted towards the end of the Nineteenth century by Paul Otlet and Henri LaFontaine in an attempt to create a universal bibliography. Until recently responsibility for the scheme belonged to the FID (Federation Internationale de Documentation), this responsibility was passed to a consortium of publishers (the UDC Consortium) in 1992. The original purpose of use for ordering and indexing entries in a printed bibliography have since been overtaken by its use for indexing and retrieval in computer based systems. The scheme consists of 60,000 classes (divisions and sub-divisions) as well as a number of auxiliary tables to describe countries, etc.

2.2.1. Extent of usage in Internet services

At least five Internet services are currently using UDC:

- *BUBL*
- *GERHARD*
- *NISS Information Gateway*
- *OMNI*
- *SOSIG*

A brief e-mail questionnaire was sent to staff at each of these services (see Appendix 1). The answers are summarised throughout this report.

BUBL

The *BUBL* Subject Tree <URL:http://www.bubl.bath.ac.uk/BUBL/home.html> aims to give comprehensive coverage of UK Internet resources in all subject areas.

The original Subject Tree uses UDC, but it should be noted that *BUBL* is in the process of transforming into a new service called LINK that will be using the Dewey Decimal Classification Scheme.

BUBL do not classify individual items, but do use the UDC to provide browsable sections for each subject area. The depth to which the classification scheme is used varies across the different subjects.

GERHARD

In the Deutsche Forchungsgemeinschaft (DFG) funded project *GERHARD* (German Harvest Automated Retrieval and Directory) <URL:http://gerhard.bis.uni-oldenburg.de/> the UDC is used in the enlarged and multilingual version of the ETH library Zürich. The aim of the project is to establish a service for searching and browsing German Internet resources. The documents are gathered by a robot, matched to the UDC entries by computer linguistic algorithms to create an searchable index and an automatically generated subject tree. The project started in October 1996 at the university library of Oldenburg. A prototype will be available during May 1997.

NISS

The *NISS Directory of Networked Resources* <URL:http://www.niss.ac.uk/subject/index.html> is a selective service that covers all subject areas.

NISS uses UDC in some detail, and browsing *NISS* involves working through UDC hierarchies, with the numbers displayed on the screen, above each section, rather like a virtual shelf mark. The Directory may be browsed in UDC number "inverted tree structure", UDC number linear structure (shelf order) or alphabetical subject heading order. All three output formats may be linked to from: <URL:http://www.niss.ac.uk/subject/index.html>.

NISS do not normally classify beyond the decimal point, although there are exceptions in the 'computing' and 'geography' sections.

Classmarks are added to the Directory on an ad hoc basis (i.e. if a sufficient number of new resources warrant the use of a new UDC classmark then one is added), since it was never intended that the full range of UDC classmarks would be used.

OMNI

OMNI (Organising Medical Networked Information) <URL:http://www.omni.ac.uk/> is a selective subject service that catalogues resources relating to medicine.

OMNI currently use UDC to create browsable sections. However, they also use a subject-based classification scheme, the NLM, which is used to create separate browsing sections. It is interesting to note that like *BUBL*, *OMNI* plan to stop using UDC in the near future. They will only use NLM.

OMNI currently classify in as much detail as possible with the NLM scheme, but find this more difficult with the UDC scheme.

SOSIG

SOSIG (The Social science Information Gateway) <URL:http://sosig.ac.uk/> is a selective subject service that catalogues resources relating to the social sciences.

SOSIG does not use the UDC in its complete form, but has drawn upon UDC social science classification numbers to create the browsing sections of the service. A selection of 26 UDC numbers are currently used for the browsing sections. In cataloguing however, a larger list is used - 57 numbers have been selected with a view to increasing the number of browsing sections when a suitable number of resources have been placed in the new sections. No other UDC numbers are currently used.

The detail of the numbers varies from being at the top of a hierarchy (e.g. Philosophy = 1) to being fairly low down the hierarchy (e.g. Environmental Issues = 551.588). One advantage of classifying Internet resources is that you can assign more than one number to a resource, since they do not need to put in numerical order on a shelf - they can be kept in two place at once.

2.2.2. Extent of usage in traditional and other online services

UDC is used in a number of online catalogues, databases and information retrieval packages. There are no current figures available for the extent of usage but a survey in 1977 found that of thirty one countries in Europe, eight used UDC in their national bibliographies with a further four countries adopting the scheme in the 1980's (McIlwaine 1991). The same paper mentioned use of the classification in Latin America and six French speaking African countries. In the UK UDC is used by a small number of university libraries including Aberdeen, Liverpool and Edinburgh. It is also used by a number of small specialised libraries such as the Scott Polar Research Institute, the Royal Greenwich Observatory and the British Architectural Library. Databases using the scheme include the *HSE-line* database on *ESA-IRS* and the *HELPIS* database on *BLAISE-LINE*. The scheme is used in the *CDS/ISIS* retrieval package and is available as a sort option in the *InMagic* information retrieval package.

2.2.3. Multilingual capability

Because the UDC is based on numerical notation the scheme is not language dependant. The classification exists in several different languages including English, French, German and Japanese.

Internet services

GERHARD was the only service surveyed that will use the multilingual capability of the UDC.

2.2.4. Strengths and weaknesses of the scheme

Strengths of UDC

UDC is an agreed international standard which means it is widely recognised, used and available. It also means it is regularly (if not frequently) maintained. As already noted the scheme is not language dependant and exists in several different languages. The structure of the classification allows composite codes to be assigned to provide complex and detailed description of the subject content of a document or resource.

Strengths as seen by Internet services

- **It is a universal standard and is widely used**

Two of the services said they originally chose to use UDC so that it would be compatible with other key Internet services being developed in the UK at the time. The fact that it was being used by *BUBL*, and

NISS were the key reasons for choosing the scheme. However, since then both *BUBL* and OMNI have decided to drop UDC despite this strength, which implies it has weaknesses which can over-ride this advantage.

- **It is flexible, and can be adapted to suit the needs of different services**

UDC is a sizeable and comprehensive classification scheme which gives it a certain amount of flexibility. A number of the services said they were able to adapt it to suit the needs of their particular service, and cited this as an advantage. *NISS* suggested that UDC could be used to a variety of different levels of precision, which suited them. *BUBL* found the top hierarchy of UDC (i.e. the first sub-divisions were suitable for the way they wanted to present their subject tree. *SOSIG* said that a suitable subject specific classification scheme could not be found for the social sciences, and UDC covered the subjects that fell within the scope of *SOSIG*.

- **It is usually free to use**

None of the services (except Gerhard) had had to pay to use UDC, and had had no problems over copyright when they published sections of it on their Web pages (there would have been copyright or licensing issues had the services used the machine-readable Master Reference File which belongs to the UDC Consortium).

Weaknesses of UDC

One of the main criticisms of the UDC is that the scheme is out of date. New knowledge is continually developing and existing knowledge is being redefined which causes problems for large schemes such as the UDC. Part of the problem of revisions to the scheme was also due to the unwieldy committee structure that was in place when the classification was under the control of the FID. The UDC Consortium have adopted a much more flexible approach to revisions of the classification, which are now done on a contractual basis.

The complex structure of the scheme is also considered a problem, the main tables of the classification can be combined with auxiliary tables and punctuation to express detailed concepts and relationships.

Many institutions and libraries only use a simplified version of the scheme (Buxton 1990). An introductory guide to the use of the UDC was published in 1993 and revised in 1995 to help users with the application of the scheme (McIlwaine 1993; McIlwaine and Buxton 1995).

Weaknesses as seen by Internet services

- **It is not updated frequently enough**

Both *SOSIG* and *NISS* suggested that UDC was not updated frequently enough and that this caused problems. *NISS* commented that the main divisions of UDC (the top level) have their roots in the 19th century and are not intuitive to modern academics. *SOSIG* said they found two problems with currency: Firstly, some of the subjects seemed to be old-fashioned and outdated. For example, UDC has a section called 'feminism' which *SOSIG* users have suggested should be called 'gender studies'. Secondly, some subject areas seem to have 'out grown' their UDC section. The UDC hierarchy does not always grow at the speed required to keep up with subject areas that have developed significantly over recent years. For example, 'environmental issues' and 'development studies' are growing areas where there is a lot of interest and a lot of new resources, but these are not particularly well catered for by UDC.

- **It is weak in certain subject areas**

Some services felt that UDC did not cater adequately for some subject areas.

- *SOSIG* found the hierarchy did not give a high enough status to Environmental Studies, Developmental Studies.

- *NISS* suggested UDC is particularly weak in Medicine and Health sciences, which was borne out by *OMNI* who point out that UDC is not generally used in medical libraries, and who have now decided to drop UDC because of its weaknesses in the medical field.

- **It is complex to use**

Three services suggested that UDC was too complicated to use (*BUBL*, *OMNI* and *BIZ/ED*).

- *OMNI* found UDC was too complicated to use in the classification of detailed and complicated subjects. UDC requires the formation of composite classification codes to express complex subjects,

whereas ROADS software wants an enumeration of all the codes you want to use. If you don't have this in electronic from this is difficult.

- *NISS* suggested that the decimal notation does not reflect the true hierarchy of subjects, which means that the hierarchical tree structure can't be maintained fully automatically, and that the implementation of wildcard searching presents problems.

2.2.5. Integration between classification scheme and other systems e.g. controlled subject headings

The *ETHICS* (ETH library Information Control System) at ETH (Eidgenössische Technische Hochschule) in Zürich link each UDC number to related thesaurus terms in English, French or German.

- ***NISS* enlarge the 'computing' section of UDC with their thesaurus**

Before the revised UDC schedules for 'computing' were published, the *NISS* service created more detailed classification under the classmark 518, the 'computing' section, by adding new numbers and headings that did not exist in the official UDC. They took the new headings from a descendant of a former *ASK/NISS* Thesaurus of software classifications. The headings from this were adapted with numeric notation to slot in under a classmark 518

2.2.6. Linking UDC to third party classification data

Bibliographic records supplied by national libraries and bibliographic utilities are able to link classification data to other classification schemes through the MARC record (see 2.1.5 and 2.1.6: DDC review).

Internet services

None of the services surveyed do this, although there is currently some potential for linking UDC and NLM classification notation in *OMNI*.

2.2.7. Digital availability

A Master Reference File of the scheme (in English) was created in 1993 and is currently maintained by the Technical Director of the UDC in the Hague. This contains approximately 60,000 classes (divisions and sub-divisions). Copies of the file are available under licence for a period of three years.

Internet services

GERHARD had multilingual and enlarged version of the scheme. *NISS* and *SOSIG* have abridged versions on their Web pages:

SOSIG

SOSIG has a selected-group of numbers from the social science sections of UDC available on their Web pages: <URL:http://sosig.ac.uk/Subjects/udc-list.html>.

SOSIG also links to the 'UDC in brief' given on the *NISS* Web server.

NISS

NISS offer a guide to 'UDC in brief' This gives more detail of the main subject headings and numbers in the scheme: <URL:http://www.niss.ac.uk/resource-description/udcbrief.html>

NISS uses UDC in some detail, and so in browsing *NISS* the UDC numbers are featured electronically on the screen.

2.2.8. Copyright

The copyright of the machine readable Master Reference File belongs to the UDC Consortium from which licenses can be purchased. Hard-copy versions of the classification are the responsibility of the individual publishers.

Internet services

Out of the Internet services surveyed, only the *GERHARD* project had to pay an annual fee of 5,000 Dutch Guldens.

2.2.9. Extensibility and development effort of the scheme

The responsibility for developing and managing the classification belongs to a Consortium, this is made up of members from Britain, Holland, Belgium, Spain and Japan (who represent the major publishers of the scheme) and the FID who were the previous owners of the classification. Developments and revisions are undertaken on a contractual basis; recent or ongoing revision work is being carried out in Astronomy, Linguistics and Philology, Medicine and Computer Science. The Consortium publish revisions annually in the publication *Extensions and Corrections to the UDC.* Other plans include moving the UDC towards a faceted classification and co-operation with the Dewey Decimal Classification Committee to jointly publish area tables so that both schemes would have a standard set of notations for expressing countries.

2.2.10. Other Issues

- **Interoperability between Internet services: mapping between schemes**

It seems that at the outset eLib funded subject services were keen to try and use the same schemes as other services with a view to being interoperable at some point in the future. Two of the services (*SOSIG* and *OMNI*) chose UDC for this reason. However, *OMNI* make the point that there is no need to use the same scheme as long as one scheme could be mapped onto another. *OMNI* find that their users much prefer the browsing sections that are based on the NLM scheme to the sections based on UDC. They plan to stop using UDC for this reason, but still see interoperability as being viable, since the subject headings from the two schemes could be mapped onto each other automatically for this purpose.

- **Classification and subject indexing in more than one scheme**

BUBL made the point that numbers from two different schemes can be simultaneously assigned. For their new service they use an OCLC CD-ROM to assign Dewey numbers and LCSH simultaneously.

2.3. Library of Congress Classification System (LCC)

One of the world's most widely spread classification schemes is the Library of Congress Classification System (LCC). This is largely due to the fact that every exported record from the Library of Congress contains their own classification of the item. Apart from being dominating, it is quite old: LCC will soon celebrate its centenary. In 1899 the Librarian of Congress Dr. Herbert Putnam and his Chief Cataloguer Charles Martel decided to start a new classification system for the collections of the Library of Congress (established 1800). Basic features were taken from Charles Ammi Cutter's Expansive Classification. LCC is an enumerative system built on 21 major classes, each class being given an arbitrary capital letter between A-Z, with 5 exceptions: I, O, W, X, Y (these appear at the second or third level in the notation for various subclasses). After this was decided, Putnam delegated the further development of different parts of the system to subject specialists, cataloguers and classifiers. Initially and intentionally the system was, and has remained, decentralised and the different classes and subclasses were published for the first time between 1899-1940. This has lead to the fact that schedules often differ very much in number and the kinds of revisions accomplished.

LCC notations are composed of repeated letters and numbers. Capital letters are, as mentioned above, used for main and subclass notations, for subdivisions further down the hierarchies LC uses Arabic numerals (i.e. Urban Transport = HE 305-311). There is no official comprehensive index to the LCC, the scheme is very extensive and included in about 46 volumes published by the Library of Congress.

The Library of Congress has developed a USMARC Format for Classification Data which allows classification data to interact with other USMARC bibliographic data and authority files (Guenther 1992; Guenther 1996). The format has been designed for use with DDC and LCC, (the two major classification systems used in the US), but permits communication with other classification schemes, especially UDC and NLM. A machine-readable version of DDC has been available for some years as part of its Editorial Support System (ESS), and recognising the benefits of this, the Library of Congress decided in 1993 to convert their 46 schedules into machine-readable form for the USMARC classification format. When complete, the LCC database will hold in the region of 450,000 classification records (Vizine-Goetz 1996a).

In this review, the use of Library of Congress Subject Headings (LCSH) in an Internet context will also be investigated to a limited extent.

2.3.1. Internet services using the Library of Congress Classification System and/or Library of Congress Subject Headings

2.3.1.1 Library of Congress Classification System (LCC)

There are several services on the Internet which claim that they use the LCC for classification of resources.

CYBERSTACKS

CyberStacks <URL:http://www.public.iastate.edu/~CYBERSTACKS/> is a centralised, integrated, and unified collection of World Wide Web and other Internet resources categorised using the Library of Congress classification scheme (McKiernan 1997). Resources are organised under one or more relevant Library of Congress class numbers and an associated publication format and subject description. The person who is responsible for the service, Gerry McKiernan of Iowa State University, has chosen to provide information within the six main classes Q (Science), R (Medicine), S (Agriculture), T (Technology), U (Military Science), V (Naval Science). The resources are first categorised, within a broad classification (i.e. Chemistry, QD), then within narrower subclasses (i.e. Physical & Theoretical Chemistry, QD 450-731), and then finally listed under a specific classification range (i.e. QD 467 Classification. Periodic Law). LC subject headings are not used and searching is not possible as the service is designed for browsing. Within the selected classes, the entire LCC notation scheme is published and as the service still is under construction, there is often a lack of content - no resources or links - behind the headings.

The WWW Virtual Library

The WWW Virtual Library
<URL:http://www.w3.org/pub/DataSources/bySubject/LibraryOfCongress.html> is a quite

comprehensive distributed subject catalogue which offers several ways to enter or browse the collected resources. One approach is through the LCC system, but this order is only applied at the very first level to produce a global ordering of a site that, in the most part is just an alphabetic list. All main classes in the LCC system are represented. The WWW Virtual Library does not use the classification notations themselves at all in their LCC structure, only their corresponding headings. The service can not be searched only browsed.

Two libraries trying to organise Internet resources in accordance with LCC

- Cardinal Stritch College Library <URL:http://www.stritch.edu/~ldiodato/cscindex.html>
- Seattle Pacific University Library- Subject Oriented Resources <URL:http://www.spu.edu:80/depts/library/second/>

Seattle Pacific writes the following about their choice:

1. "The Library Web page is in LC order so that you can easily find additional sources of information in the same areas where that subject is found in the collection. For instance, if you are doing research on Education, you will find both books and Web references in L." (Seattle Pacific University Library 1996)

2. "At SPU, we decided to change from Dewey Decimal to the Library of Congress, because it provides for better subject access to the collection. LC has many more subjects than Dewey, which tends to put things in broad categories." (Seattle Pacific University Library 1996))

Both libraries only use the classification system as a way to organise resources at a first broad level (i.e. L Education, HA Statistics) and the classification notation disappears at the second level. All (main and sub-) classes are meant to be covered but this is done quite poorly and in a haphazard way. LC subject headings are not used and the sites are not searchable.

T.F. Mills Home Page - Some Humanities Links in Library of Congress Class Order

T.F. Mills <URL:http://www.du.edu/~tomills/> has organised his favourite sites according to classes in LCC on his private home page. This is done merely on first level and the subjects are chosen at random. It is not possible to search the site. He does not use LC subject headings.

2.3.1.2 Library of Congress Subject Headings (LCSH)

NetFirst

NetFirst <URL:http://www.netfirst.ac.uk/> is a commercial search service run by OCLC composed of a growing collection with approximately 60,000 full bibliographic descriptions of Internet resources. The resources are compiled by OCLC staff and classified in Dewey Decimal Classification (DDC) and the opportunity is given to search for DDC notations. Until recently browsing has only been provided in indexes arranged alphabetically, but a browsing structure based on the DDC notations for each record has recently been made available (Oehler 1996). The service also uses LC Subject Headings to make the resources searchable (see also 2.1.1).

***NISS*: National Information Services and Systems**

NISS <URL:http://www.niss.ac.uk/> is a service classifying Internet resources and offering several ways of access to them, through both searching or browsing. All resources in *NISS* are described by records in a Resource Descriptions Database (which underlies the Information Gateway). The records contain, among other information, UDC classifications and these constitute the basis for the browsing structure, and they are searchable as well. The goal is to cover all subjects.

NISS now wants people to give (four) LC subject headings when sending in a resource through the 'send-in-a-resource' form, but it remains optional to add that information. On their information page they write: "these headings represent a structured natural language thesaurus which has been applied to all records in the NetFirst service ... In order to make *NISS* resource descriptions more compatible with NetFirst records we are encouraging the use of Library of Congress Subject Headings as a supplement to the UDC classification ..." (*NISS* 1996) (see also section 2.2.1. on UDC).

INFOMINE

INFOMINE <URL:http://lib-www.ucr.edu/> started January 1994 as a project at the Library of the University of California, Riverside. It "currently enjoys participation from librarians at all nine

University of California campuses and Stanford University and is a good example of a multiple-campus, shared Internet resource collection project" (*INFOMINE*, 1997).

It is a combined search and browse service and it offers descriptions of and access to about eight thousand Internet resources, all of which are said to be of academic interest. Browsing can be done by date-'what's new', title of resource, table of contents (i.e. resource titles arranged under their subjects), subject and keyword. Searching is done in the following fields: title, subject and keyword. INFOMINE uses LCSH when cataloguing Internet resources, so each record in the database is given a number of Library of Congress Subject Headings, as is normally done for books in libraries. These subject headings are then used as subject-terms under which the resources are organised in the browsing structure, only they are displayed alphabetically. Before the user gets the opportunity to search or browse he/she has to choose one of the following main subjects/classes on the top page of the service: 1. Biological, Agricultural, and Medical Sciences, 2. Government Info., 3. Instructional resources: K-12, 4. Instructional resources: University , 5. Internet Enabling Tools, 6. Maps & GIS, 7. Physical Sciences, Engineering, Computing, and Math, 8. Regional & General Interests, 9. Social Sciences and Humanities, 10. Visual and Performing Arts

Each of these are called a *INFOMINE* and they not related to either LCC or LCSH practice. When the search results are presented there is a list at the end where the same subject heading can be searched for in the other *INFOMINE*s.

The minimum number of LCSH applied to each resource are 2-3 but many records have more than this, usually about 4-8 headings in addition to several key-words. There is no maximum upper limit on the number of LCSH terms used <URL:http://lib-www.ucr.edu/pubs/postlcsh.html>.

General

Several of the examples presented here are trying to cover a wide range of subjects, "the total knowledge of the world"; all (or most) of the classes in LCC. Not one of them is aiming at a limited service; only covering one subject. LCC is an international universal classification scheme, and is therefore unlikely to be the best choice for a service providing extensive information within one specific subject area. For that it would probably be more convenient to use an international subject-specific scheme. However, *CyberStacks* covers Science and Technology and seems to find LCC detailed enough for these subjects.

There is a difference between *NetFirst* and *NISS* and the other services mentioned. The former are making database records of each resource which then are organised mechanically in classification systems/structures, whereas the other services are making systems of Web pages, corresponding these to the classification scheme and linking to the resources from the appropriate schedule.

The classification in these Web page based services is mostly produced in a rather superficial way and subject headings are not used to describe the resources, so is there no opportunity to build connections between subject headings and classification. The notations exist all the way through different levels in *NetFirst* and *NISS*, whereas they disappear rapidly in the other services, with the notable exception of *CyberStacks.* The possibility to *search* for classification numbers (or subject headings) is only offered in *NISS* and *NetFirst*.. To navigate in the other services the user can only browse. Some of the services offer an alternative access to their collections by listing the titles of the resources alphabetically.

Often the services seem to have changed the original heading of the notation, and several of them have not used the precise notation belonging to LCC. As soon as this takes place, the services have altered the schemes, and to all intents and purposes, have developed a different classification scheme.

2.3.2. Extent of usage in traditional and other online services

LCC is used extensively in the United States, Canada and Australia, principally in academic libraries in both card and online catalogues. LCC notations are present in many records in OCLC (Online Computer Library Centre) and RLIN (Research Libraries Information Network), but their presence depends on the institution inputting the bibliographic record. The Library of Congress, of course, always provides an LC notation, but other libraries have the option of supplying other classification numbers or none at all.

This can be compared with the use of LCC and LCSH in the Nordic countries. Out of the 40 largest libraries in Sweden one uses LCC, and another uses a local version of LCC. LCSH are not used at all in Sweden whereas in Norway the subject headings are used to some extent but LCC is not used at all.

Many library catalogues in the world have hidden LCC notations in their records since the cataloguers deliberately do not always erase this field when importing LC records mechanically. One example is LIBRIS, the Union catalogue of Swedish research and special libraries.

Some OPAC systems can search on Library of Congress Classification numbers. It is contained in field 050 of the USMARC bibliographic record.

2.3.3. Multilingual capability

LCC is an American system and has no multilingual capability. There is no well-known translation of the LCC schedules. The notation itself is not language dependent since it is a enumerative system, using letters (Latin) and numbers (Arabic) that are used in a considerable part of the world. Some classification numbers have captions in multiple languages, but these are primarily in the law schedules.

2.3.4. Strengths/weaknesses in the subject areas of interest for the DESIRE testbeds

Since LCC is a general classification scheme it is likely to be less detailed in specific subjects than subject specific schemes. There are no examples of services using LCC for the classification of resources within a single subject area, like subject based information gateways.

Engineering

This class is up-dated regularly and thus new subjects or aspects are added quite quickly. The fairly new and fast growing subject 'Computer Engineering, Computer Hardware' is covered by the notation span TK 7885-7895, a subclass to 'Electrical Engineering, Electronics, Nuclear Engineering'TK 1-9971.

Being a general scheme, LCC is not as good as Ei (Engineering Information Inc.'s subject specific classification scheme) and other subject specific classification schemes for detailed classification of large collections of engineering resources.

Art

The class Fine Arts (N) still has many empty notations. It is organised by form (sculpture, painting etc.) first, then by chronology or nationality, and finally by artist.

Social Sciences

The coverage of the social sciences in LCC is covered in different ways depending on which country is being classified, i.e. sometimes Law and Economics are covered, sometimes they are not. LCC reflects an American way of understanding the social sciences and how to organise resources within the subject. This does not always fit the ideas or demands of other countries.

The social science subject service *SOSIG* chose UDC to be compatible with other national services in UK such as *NISS* and *BUBL*. They did not take into account any other universal schemes such as LCC or DDC at the time that they made this choice.

General

Marcella and Newton note that "LC is the least international of the major general classification schemes. In its coverage it predominantly reflects a national collection; there is a distinct bias towards the social structure, history, law and cultural concerns of the United States. The notation is complex and not truly comprehensible internationally. In particular, the use of Cutter numbers, which has a linguistic dimension, is not likely to be consistently applied internationally." (Marcella and Newton 1994).

By contrast, the U.S. bias among LCSH in recent years has diminished, as a consequence of the inclusion of headings contributed by libraries other than the Library of Congress.

LCC's development history means that it can be seen more as gathering of a whole range of special subject classification schemes. It is up-dated in subject areas that change regularly but special libraries still do not seem to think that their subjects are covered well enough. The system is developed through continuous revision, at least in part as there are schedules that have not been up-dated since the nineteen-sixties, i.e. PB-PH Modern European Languages (1966).

LCC is sizeable and comprehensive and there are hundreds of letter-number combinations left making it suitable for expansion in the future.

2.3.5. Integration between classification scheme and other systems e.g. controlled subject headings

The authority record for subject headings has a field for a classification number if there is a correlation. The USMARC bibliographic record format has fields for both classification data and for controlled subject headings. This provides one mechanism for linking the two.

The USMARC Format for Classification Data has fields for Index Terms (fields 700-754), so that LCC (or DDC) classifications can be expressly linked with subject or thesaurus terms like those in LCSH or MeSH (Guenther 1996, p. 190).

Vizine-Goetz (1996b) notes that "For LCC, explicit links between LC subject heading and class numbers occur in LC Subject Authority records that contain classification number fields. In an analysis of the LC Subject Authority file, Vizine-Goetz and Markey (1989) found that about 43% of topical subject heading records (MARC tag 150) contain LC classification number fields. Science and technology classes account for almost half (47.72%) of the class numbers. Efforts to improve the index to LCC may also lead to better links between LCC and LCSH."

"… the LC cataloging and Policy Support of Office is reviewing the index structure of the LCC schedules and is consulting with classification expert Lois Chan on the design of a combined index to LCC. It is very likely that this work could lead to future efforts to form better links between LCC and LCSH." (Vizine-Goetz, 1996b)

The printed LCSH does provide examples of possible LC classes, but these are only suggestions. For some subject headings a range of LC classes are given, in others several different ones depending on the aspect of the subject matter, and for many there are no classes mentioned at all.

LCC is very extensive and unfortunately there is no official comprehensive index. The printed edition of the Library of Congress Subject Headings is the publication most likely to serve as a substitute for such a comprehensive index. It contains references to one or more class numbers after entries and subdivisions but no effort is made to maintain class notations in LCSH.

2.3.6. Linking LCC to third party classification data

LCC is linked to DDC and other classification schemes (including UDC and NLM) in many MARC catalogue records supplied by bibliographic utilities (for more details see also 2.1.5 and 2.1.6: DDC review).

2.3.7. Digital availability

With reference to the USMARC Format for Classification Data, LC made a commitment in 1993 to complete the conversion of all LCC schedules into machine-readable form (Guenther 1996, p. 178) and this classification database is still growing and being improved. This machine-readable version of LCC fulfils a need first identified by Chan (1986). Currently, only a test file is available: <URL:http://lcweb.loc.gov/cds/newform#lccr>

Many LCC schedules, though not all, are available on a CD-ROM product called *Classification Plus*, available from the Library of Congress. Information on this product is available at: <URL:http://lcweb.loc.gov/cds/cdroms1.html#classplus>

LC does not offer any schedules or manuals online, i.e. on the Internet, but LC is represented on the Net and their Cataloging Distribution Service (CDS) describes their product range at: <URL:http://lcweb.loc.gov/cds/cdsintro.html>

There is, however, a private initiative by Matt T. Rosenberg, who has published LCC at a page called 'Library of Congress Classification System' at: <URL:http://www.geocities.com/Athens/8459/lc.html>. This site comprises LCC in hypertextual form and it allows browsing of notation ranges on the World Wide Web. It includes only schedules and does not include resources classified in accordance to the system. The site is not an authorised official Library of Congress service and as it is still under construction, not all classes are currently completed.

2.3.8. Copyright

The classification system is not copyright in the United States but it is elsewhere. As with DDC (see section 2.1.8) the classification numbers would be able to be used without restraint but not the accompanying textual material.

More information on LCC can be obtained from the Cataloging Distribution Service of the Library of Congress. <cdsinfo@mail.loc.gov>

2.3.9. Extensibility and development effort of the scheme

Extensibility

1. LCC is an enumerative system which could make it hard to adjust to new developments within subjects and to totally new subjects but there are "hundreds of number-letter combinations compatible with the notation that have not yet been employed or have even been retired in favour of new locations. ... The scheme will to accommodate for a long time the many new subjects and aspects of subjects not yet anticipated." (Wynar 1992, p. 351)
2. After the initial numbers and letters, another set of mixed letters and numbers, called Cutter numbers, an adjusted version of the "Cutter author-mark idea", are used as a mean to achieve different alphabetic sub-arrangements. When decimals are used this do not signal subordination but that new classes/subclasses has been added to sections of the scheme where there were no further integers available. Thus the decimals are used to "allow a new topic or aspect to be inserted into an established context." (Wynar 1992, p. 369)
3. "Geographical and chronological subdivisions are often relegated to auxiliary tables. Frequently the two concepts are combined... But other principles of division may also be found in tabular form." "A few auxiliary tables, especially certain geographic lists from the H schedule, are said to "float". That is, they appear, usually in slightly variant forms, in other schedules. Thus, the Table of Countries in One Alphabet from the H schedule shows up in mutation as an auxiliary table for schedules C, D, E-F, T, U, and V." There are two different kinds of auxiliary tables, simple and compound. Simple implies that the tables "carries only one sequence of numbers that can be interpolated directly into the corresponding number spans in the schedule. Whereas the compound tables "supply more than one sequence for the same list of subtopics"(Wynar 1992, p. 365). Some of LCC auxiliary tables are designed for use with more than one specific class notation or span, others apply only to one subclass.

The LCC are generally not able to reflect the language of the material being catalogued, with some exceptions in some areas of the literature schedules (Wynar 1992).

Development effort

Up-dating is accomplished in committee and published in a number of publications (both printed and online), most of which are available from the Library of Congress. Revised editions of individual schedules, *Library of Congress Classification - Additions and Changes*, *Library of Congress Classification Schedules: A Cumulation of Additions and Changes*, *Cataloging Service Bulletin*, *Library of Congress Subject Headings*. Information about developments and news in LCSH can also be found on the Internet. The *LC Weekly Subject Headings List* is at <URL:gopher://marvel.loc.gov:70/11/services/cataloging/weekly>. LCC is said to be flexible and "functionally up-to-date" since revisions are made independently within special classes and subclasses.

2.3.10. Other issues

There are concordances from the Dutch national scheme NBC to and from LCC. These concordance tables are used in the Pica automated system.

2.4. National general schemes

2.4.1. Nederlandse Basisclassificatie (BC)

The Nederlandse Basisclassificatie (Dutch Basic Classification) is a national scheme designed for use within the Shared Cataloguing System of Pica. Pica, the Dutch Centre for Library Automation, is a non-profit organisation providing systems and services for the majority of Dutch academic and public libraries and for a number of library networks in Germany (Die Deutsche Bibliothek, Gemeinsamer Bibliotheksverbund, and Hessisches Bibliotheksinformationssystem). The BC has two objectives: The first is to make shared subject indexing possible by the use of the same classification scheme by all Pica libraries. Secondly, to enable co-ordination of collection development in the different libraries in a project based on the Conspectus system developed by the Research Libraries Group (RLG) in the USA.

The BC consists of 48 main hierarchies grouped in five clusters (General; Humanities; Sciences; Engineering; Social sciences).

2.4.1.1. Usage

The BC is a relatively new scheme, developed in the period 1986-1991, for use in an OPAC.

Circa 200 libraries use Pica's Shared Cataloguing System (GGC), but only 18 of them actually add BC notations to their records. Among those 18 are the most important academic (university) libraries, which means that in fact a very high percentage of the records in the GGC is covered. The BC was translated in German for use by a number of German libraries which adopted the Pica cataloguing system (Facharbeitsgruppe Sacherschliessung 1995).

There is some unofficial use of the BC outside the Pica system. The publisher's database NESTOR uses the BC for recently published books, and it is used by the Rijksdienst Beeldende Kunst for the classification of Dutch posters.

Use of the BC in NBW

In *NBW* (Nederlandse Basisclassificatie Web) <URL:http://www.konbib.nl/basisclas/basisclas.html> the BC is used for the classification of Internet resources. *NBW* is a subject service of quality resources, maintained by the Koninklijke Bibliotheek (KB) in co-operation with a number of Dutch academic libraries. Each resource is given a simple catalogue entry containing title, URL, descriptive note in English and selection signature (identifying the library and the subject specialist who selected the resource). The descriptions of the resources are searchable via a WAIS-index. The BC notation is added to the record of the resource with the corresponding heading in English and in Dutch so it is possible to search the BC on both English and Dutch BC corresponding headings, while the descriptive notes are only available in English. In *NBW* the BC serves also as a navigation tool for browsing the descriptions of resources.

2.4.1.2. Multilingual capability

As a numerical notation the BC is not dependent on language. The corresponding headings of the 'original' BC are in Dutch. At the present time translations in English and German are available.

2.4.1.3. Strengths and weaknesses

Division in the BC of the disciplines the DESIRE test-beds focus on:

***Art*:** divided in 2 main hierarchies: art (20) and art forms (21).

Engineering: 9 main hierarchies: technical sciences (50); materials science (51); mechanical engineering (52); electro-technology (53); computer science (54); traffic and transport technology (55); civil engineering (56); mining engineering (57); and process technology (58).

Social Sciences: divided in 14 main hierarchies: social sciences (70); sociology (71); cultural anthropology (73); geography (74); recreation, leisure (76); psychology (77); welfare, social assistance, theory of adult education, personnel management (79); pedagogics (80); education (81); economics (83); business administration and management, organisational science (85); law (86); science of social and public administration (88); political science (89).

Advantages

Having been developed quite recently the division of disciplines is in accordance with modern developments. For instance, there are separate hierarchies for environmental science and computer science.

Disadvantages

- Some hierarchies are subdivided along geographical lines, like history and art history, which are not always the most appropriate subdivision.
- The BC tends to be unbalanced: not all the hierarchies are subdivided with the same level of specificity. Some hierarchies are subdivided in great depth, others in very broad classes. Examples: 'language and literature' have only one to four classes per language area, depending on the relative importance of the language (from a 'Eurocentric' point of view). 'History of Europe' is only one class, while 'codicology' is subdivided in ten classes.
- Because the BC is meant to classify according to discipline, interdisciplinary subjects are very difficult to classify. The same applies for instance to resources that concern a certain country, its history, language, art, culture etc. Those do not fit in one BC-class.
- Some specialisations are a subdivision of more than one discipline, so for instance 'philosophy of religion' is a class of 08. philosophy as well as of 11. theology.
- 'Education' is divided in two categories: 'pedagogics' and '(school) education'.
- Every hierarchy has a class 'general' and a class 'other'; the difference is not always clear.
- Because all the classes have a notation of four digits, with a dot after the first two, the relative place of a class in the hierarchy is not always clear, as the hierarchy may comprise more than two levels.

2.4.1.4. Integration between classification scheme and controlled subject headings

The Nederlandse Basisclassificatie is a numeric root-classification. Classification of the main content (on title level) indicates the discipline the subject belongs to. For more specific subject indexing an additional thesaurus was developed, derived from the thesaurus used by the University of Amsterdam. This GTT (GOO Trefwoorden Thesaurus = Shared Thesaurus of Subject Headings) consists of: general subject headings, geographical descriptors, corporations as subject, titles/name of works (of art), form descriptors and genre descriptors. There is no pre-established relation between BC notation and GOO subject heading, a subject heading can be combined with any BC notation. The subject headings form an open system, maintained in the GOO thesaurus, part of the Shared Cataloguing System, with a record for each subject heading. There is a separate thesaurus for names of persons. The classification scheme (BC), subject headings (GTT) and, if applicable, a historical period, together constitute what is called the *Gemeenschappelijke Onderwerpsontsluiting* (GOO), or Shared Subject Indexing.

The GOO-subject headings, which in Pica are used together with the BC to allow more specific indexing are not used in NBW, because it would present problems to use the subject headings outside the Pica system. For instance it is doubtful whether new subject headings can be added that are not used in the Cataloguing System, but (only) in the Internet environment. Furthermore participation in NBW by non-Pica libraries would pose a problem, because the GTT is incorporated in the Pica system. Instead of using the subject headings the intention is to extend indexing and retrieval facilities of the NBW in other ways, following new developments in indexing on the Internet.

2.4.1.5. Linking to third party classification data

There is no linking to other classification data.

2.4.1.6. Publication in digital form

The online BC-thesaurus is part of the Pica System, with a record for every class. The first printed edition of the BC appeared in 1989, an amended working edition in 1991, and a second edition in 1992.

The Nederlandse Basisclassificatie is also available on the Internet:

Dutch version: <URL:http://www.konbib.nl/basisclas/bcned.htm>

English version: <URL:http://www.konbib.nl/basisclas/bc.htm>

2.4.1.7. Copyright issues

The owner of the copyright for the printed edition of the BC is the *Stuurgroep GOO* (Steering Committee GOO). The BC may be used freely outside the Pica System; no costs are involved. Use of the GOO-thesaurus is much more tied up with participation in the Pica System. For a search of the c. 40,000 subject headings access to the Pica cataloguing system is necessary. Mutations and additions can only be made by Pica participants and are subject to formal approval by the *Beheerscommissie GOO* (Maintenance Committee).

2.4.1.8. Extensibility and development effort provided by the authoritative body that controls the scheme

Main hierarchies consist of a two digit number, so that classes 00 to 99 are available. Only 48 of them are actually allocated. Subdivisions consist of another two digits, added to the main hierarchy and separated by a dot, so 08 is 'philosophy'; 08.23 is 'renaissance philosophy.' Each main hierarchy can be divided in 100 classes. The number of allocated subclasses is different for every main class.

The used and unused classes are distributed unevenly throughout the scheme. The following main classes are currently in use:

01 general works; 02 science and culture in general; 05 communication science; 06 documentary information; 08 philosophy; 10 humanities 11 theology; 15 history; 17 linguistics, literature 18 languages and literature per language area; 20 art sciences; 21 art forms; 24 theatre and music; 30 exact sciences; 31 mathematics; 33 physics; 35 chemistry; 38 earth sciences; 39 astronomy; 42 biology; 43 environmental science; 44 medicine; 45 veterinary medicine; 48 agricultural science; 49 domestic science; 50 engineering; 51 materials science; 52 mechanical engineering; 53 electrotechnology; 54 computer science; 55 traffic and transport technology; 56 civil engineering; 57 mining engineering; 58 process technology; 70 social sciences; 71 sociology; 73 cultural anthropology; 74 geography; 76 recreation, leisure; 77 psychology; 79 welfare, social assistance, theory of adult education, personnel management; 80 pedagogics; 81 education; 83 economics; 85 business administration and management, organisational science; 86 law; 88 science of social and public administration; 89 political science.

The body that controls the BC is the *Beheerscommissie GOO* (Maintenance Committee). The last edition of the BC (1992) has been fixed for a period of five years, expiring this year. At this moment proposals for revision may be sent to the *Beheerscommissie*. These proposals will be developed by committees of subject specialists in a given field.

When new classes are created, the first two positions of the code have to correspond with the German version, with the exception of 15 (history) and 86 (law). A new edition of the BC will follow, this new edition will again be 'definite' (not revisable) for a period of five years.

For *NBW* subdivisions of classes with to little specificity (e.g. 76.12) are considered, that won't necessarily follow these 'official' developments.

2.4.1.9 Possibilities for conversion

A conversion was developed by Pica for titles that are delivered on tape to Pica by the Library of Congress. The LCC-classification code is automatically translated into a BC notation when the new titles are read into the Pica cataloguing system.

2.1.1.10 Literature on BC

Nederlandse Basisclassificatie (1992),

Facharbeitsgruppe Sacherschliessung (1995)

Nederlandse Basisclassificatie Web (NBW). <URL:http://www.konbib.nl/basisclas/basisclas.html>

2.4.2. Sveriges Allmänna Biblioteksförening (SAB) Classification System

This is only a short review of the SAB system, for comparison with the broadly similar BC.

The SAB Classification System was published in 1921 and has been thoroughly modified twice since then, in 1956 and in 1984. The system is alphabetic and built on 25 main classes which are given the letters between A - Y (except W) and the additional Swedish letter Ä. Subdivisions are created through further addition of letters or combinations of letters. Many of the classes that could have been ordered hierarchically have, for various reasons (chiefly to avoid too long notations), been co-ordinated. Some of the classes are organised analogously, e.g. the principle of division for countries is the scheme for Geography (N) which is applied to the classes Archaeology (J), History (K) and Ethnography, Ethnology, Social Anthropology (M). The system has a few auxiliary tables that express different aspects of a subject. Some of these are general, used in many classes, others are specific for each main class.

2.4.2.1 Usage

Projects on the Internet using SAB to classify resources:

Länkskafferiet, the Link Larder <URL:http://www.ub2.lu.se/skolverket/sab_top.html>

This service offers access to 1,022 (in January 1997) quality-assessed Internet resources. The classification is used to organise the browsing structure of the service in accordance with the classification scheme, which is done automatically. The notations are not shown to the user but lie hidden in every record which also contains subject headings. Notations are not searchable but the subject headings are. All main classes are covered but as in other Internet services not all subclasses are represented.

Internetkontakt <URL: http://www.btj.se/btj/saburl/saburl.html>

Single page with links listed after the principles of SAB. Not searchable in any way and subject headings are not used at all. All main classes are covered but are applied at the second level selection very haphazardly and many subclasses in the system are not covered at all.

Systematisk internetkatalog <URL: http://www.molndal.se/bibl/subject.htm>

Pages with links listed after the principles of SAB. All main classes are represented but at the second level the choice of resources is made randomly.

Informationskällor ordnade enligt Klassifikationssystem för svenska bibliotek (SAB) <URL:http://chaplin.bibl.liu.se/sab/huvtswe.htm>

A service from Linköpings Universitetsbibliotek listing Internet links by SAB. All main classes are represented, although there are not links to all of them yet.

The latter three services only classify to the second level (i.e. Cj (KRISTENDOM) Christianity) and both show the notations to that level. From there the links are made to the sites. In the *Link Larder* though, both classification and notations are used as far down the schemes as possible but the notations are never shown to the user, not even on the very first level.

Browsing down the scheme structure takes the user as far as 6 steps along the scheme, e.g. <URL:http://www.ub2.lu.se/skolverket/sab/Koafh_a_54.html> (the URL is the only way for the user to see the classification).

None of the three services have included the whole scheme, that is they do not have SAB headings without links behind them.

Usage in traditional library catalogues

The SAB system is used by almost all public libraries and most of the university libraries in Sweden. SAB is especially predominant amongst libraries specialising in the humanities and social sciences. Research libraries in the Engineering area do not use the scheme as much as those subjects are not covered well enough.

2.4.2.2 Strengths and Weaknesses

Disadvantages

- It is a general scheme and there are not many notations left over for the development of new classes. For instance, a new subject will not be close to other related subjects but will be squeezed into the scheme where notations are still available. This makes the scheme inconvenient for the classification of Internet resources, because subjects about which a lot of information is to be found on the Internet, e.g. computers, networks, telecommunications, etc. are poorly developed in SAB.
- As was mentioned above, many of the classes that could have been ordered hierarchically, have been co-ordinated. Thus, the system is not logically ordered which makes it impossible to use for automatic classification. A computer will not understand how to organise the subclasses Fi (Italian), Fj (French), under Fh (Roman languages) since their notations are at the same level as the - regarding content - superior Fh.

2.4.2.3 Linking to third party classification data

There is no linking to other classification data.

2.4.2.4 Publication in digital form

The classification system is not available in digital form but people having access to the Swedish national catalogue for public libraries, BURK, can get to a list with the subject headings.

2.4.2.5 Copyright issues

The company BTJ (Bibliotekstjänst) owns the copyright for the SAB Classification System.

2.4.2.6 Extensibility and development effort provided by the authoritative body that controls the scheme

To find newly added subject headings and other changes in the list there is a page at <URL:http://www.kb.se/nb/AOLISTA.htm> hosted by the Kungl. biblioteket, Sveriges nationalbibliotek in Stockholm.

2.4.2.7 Literature on SAB

Klassifikationssystem för svenska bibliotek, 1984.

2.5. International subject specific schemes

A list of classification schemes and controlled vocabularies used in existing Internet services can be found in McKiernan (1996). Those in use include a number of internationally used subject-specific schemes like the National Library of Medicine (NLM) scheme and Ei, and national schemes like the Danish Veterinary and Agricultural Library Classification. Those considered here will be those relevant to the DESIRE test-bed services.

2.5.1. Art

For object description in art an important instrument is the AAT (Art & Architecture Thesaurus), but as the AAT is a thesaurus, and not a classification scheme, it isn't reviewed in this report. As more and more images become available in digital form the need arises for a classification scheme developed especially for subject description of visual images. Iconclass is such a scheme.

2.5.1.1 Iconclass

Iconclass <URL:http://iconclass.let.ruu.nl/>is an iconographic classification system, developed by Henri van de Waal (1910-1972), Professor of Art History at the University of Leiden. Iconclass is an alpha-numerical classification, hierarchically and systematically ordered, of the subjects of Western art, offering definitions and keywords in English. With this scheme it is possible to describe objects, persons, events, situations and abstractions that appear in visual images.

It is divided into 9 main categories: 1. Religion and Magic; 2. Nature; 3 Human Being, Man in General; 4. Society, Civilisation, Culture; 5. Abstract Ideas and Concepts; 6. History; 7. Bible; 8. Literature; 9. Classical Mythology and Ancient History. The code gets longer when the concept becomes more specific. Retrieval is not only possible via the alpha-numerical notations, but also via a subject index (controlled vocabulary).

Usage

An overview of institutions and projects using Iconclass can be found at: <URL:http://iconclass.let.ruu.nl/texts/institut.htm>

A CD-ROM with images of Dutch printers devices from the period 1540-1700 was produced as a pilot project for the use of the computer version of Iconclass (1992).

Multilingual capability

Although the system is published in English, the alpha-numerical notations make it language independent.

Availability

The Iconclass System and Bibliography, together with a very extensive alphabetical index, were published in 17 volumes by the Koninklijke Nederlandse Akademie van Wetenschappen (KNAW), in the years between 1973 and 1985. In the years 1990-1991, a computerised version of the Iconclass System was prepared at the Department of Computers & Humanities of Utrecht University. This version was baptised the Iconclass Browser and published in 1992 by the Iconclass Research & Development Group. It runs on computers installed with Microsoft Windows. A second edition, now including the electronic Iconclass Bibliography, appeared in 1994. The Iconclass home page also offers access to a WWW version of the Iconclass Browser which gives a good impression of Iconclass's contents, though its interface is far less powerful than the interface of the Iconclass Browser for Windows.

Copyright

Copyright for the ICONCLASS System is held by the: Koninklijke Nederlandse Akademie van Wetenschappen, Amsterdam.

Copyright for the ICONCLASS Browser: is held by the ICONCLASS Research & Development Group (IRDG), Universities of Utrecht and Leiden; Vakgroep Computer & Letteren, Utrecht University.

Copyright for the ICONCLASS Help system is held by the: Vakgroep Computer & Letteren, Utrecht University.

Development effort

The Iconclass Research & Development Group (IRDG) is the distributor of the Iconclass Browser, and the central editing board. It monitors the consistency of the scheme and decides about changes in new editions.

Other issues

Iconclass describes individual images and does not apply to collections of images.

Literature on Iconclass

Brandhorst and Van Huisstede (1992)

Grund (1993)

Iconclass home page: <URL:http://iconclass.let.ruu.nl/>

2.5.2. Social Sciences

There are no major classification schemes for the social sciences. In many cases LC Subject Headings are used, or the ERIC thesaurus. For classification, universal schemes like DDC, UDC, or LCC have been used.

2.5.3. Medicine

2.5.3.1 NLM: National Library of Medicine

The National Library of Medicine Classification covers the field of medicine and related sciences. It is a broad classification, intended to be used for the shelf arrangement of all library materials. The NLM Classification is a system of mixed notation fashioned after the Library of Congress Classification (LCC) where alphabetical letters which denote broad subject categories are further subdivided by numbers. The NLM Classification utilises schedules QS-QZ and W- WZ, permanently excluded from the LCC schedules and is intended to be used with the LCC schedules which supplement the NLM Classification for subjects bordering on medicine and for general reference materials. The LC schedules for Human Anatomy (QM), Microbiology (QR), and Medicine (R) are not used at all by the National Library of Medicine since they overlap the NLM Classification.

The headings are interpreted broadly and include the physiological system, the speciality or specialities connected with them, the regions of the body chiefly concerned and subordinate related fields. The Classification is hierarchical, and within each schedule, division by organ usually has priority. Each main schedule, as well as some sections within a schedule, begins with a group of form numbers ranging generally from 1-49 which are used to classify materials by publication type, e.g., dictionaries, atlases, laboratory manuals, etc.

Since the NLM Classification scheme is designed as a broad classification, it can be used for both large and small collections and it can also be adapted to handle specialised collections. For example, a dental library may want to expand the WU (Dentistry) schedule to meet their specific needs.

Internet Usage

The NLM is in use by the *OMNI* (Organising Medical Networked Information) service <URL:http://www.omni.ac.uk/>. *OMNI* also uses UDC but will stop using this in the near future because it is considered weak in the medical field (see section 2.2.1: UDC review). *OMNI* offers the possibility to browse the NLM in classified order or alphabetically.

Multilingual capability

The scheme is available in English, Japanese, French and Spanish.

Integration between classification scheme and controlled subject headings

Terms used in the NLM Classification: schedule headings, subheadings and class number captions are compatible with Medical Subject Headings (MeSH) descriptors. Most of the index entries in the fifth edition of NLM are in the current forms of MeSH descriptors. Others are in non-MeSH forms when no appropriate MeSH term is available to express a concept.

Availability

The fifth edition of the National Library of Medicine Classification, 1994 can be ordered from the Superintendent of Documents, U.S. Government Printing Office. The classification is not available in electronic format. When it becomes available electronically in the future, they anticipate that a normal charge of licensing fee will apply.

Copyright

Copyright belongs to the U.S. National Library of Medicine.

Development effort

The NLM Cataloguing Section has an in-house online classification file which is amenable for continuous updating in order to add new index records, new classification numbers as well as to modify existing records. As a rule, as new MeSH descriptors are added or existing ones changed, they try also to incorporate pertinent additions and changes in the classification scheme.

The fifth edition, published in 1994, contains close to 4,000 classification numbers and a comprehensive index containing over 18,000 index terms including both the index entries and cross references. Approximately 300 new classification numbers were added including form numbers which are repeated in applicable schedules across the entire classification scheme. Future changes and additions to the NLM Classification, will be announced in NLM Technical Bulletins.

Literature on NLM

U.S. National Library of Medicine (1996)

2.5.4. Science and engineering

2.5.4.1. Engineering Information (Ei) Classification Codes

The *Ei Classification Codes* are a classification scheme developed by Engineering Information, Inc. Engineering Information, also known as Ei, <URL:http://www.ei.org/> was created in 1884, at Washington University, St. Louis. Ei's aim is to identify, organise and facilitate easy access to the published engineering literature of the world.

The system has been further subdivided (1993) and now comprises six main categories, subdivided into 38 subject series and over 800 individual classes. Up to four levels of increasing specificity are provided below the main categories. It is a numeric scheme, but not hierarchical in content, but enumerative, and with some severe logical shortcomings in its structure.

Usage

Ei publishes the most comprehensive interdisciplinary engineering databases in the world, of which Ei Compendex is the most popular, with over three million records of journal articles, technical reports, conference papers and proceedings in electronic form, dating from 1970 onwards. It is also accessible on the WWW as a fee-based service.

Ei classification codes are used by two Internet subject services: *EELS* (Engineering Electronic Library, Sweden) <URL:http://www.ub2.lu.se/eel/> and *EEVL* (Edinburgh Engineering Virtual Library) <URL:http://eevl.icbl.hw.ac.uk/>.

EELS is a co-operative project of The Swedish Univ. of Technology Libraries, a consortium of the six most important research libraries in Sweden in technological subjects with some co-operation from other Nordic countries. It is developed and maintained at Lund University Library, NetLab, since the beginning of 1994 as one of the first selective subject services on the Internet using a library classification scheme and thesaurus.

EELS is arranged according to the Ei subject classification for most of subjects. The following subject fields are included: civil engineering; mechanical engineering; electrical engineering; computing; chemical engineering/chemistry; mathematics; physics; environmental engineering/science; engineering management. Some areas of interest which are not covered explicitly in the classification from Ei are for the moment maintained outside the classification, i.e. polar research and cold region technology. Descriptors (DE) from the Ei Thesaurus have been added to each resource title. Additional controlled terms, describing document types and similar are added in the annotation field of every record, since the Ei Thesaurus does not cover digital and Internet documents. A robot-generated browsable and searchable database of "all" engineering WWW pages in the Internet has been added recently. It will

soon be structured according to the Ei classification as well and will, in different ways, be integrated into the main quality controlled service.

The Edinburgh Engineering Virtual Library (*EEVL*) project started in August 1995, funded by the Joint Information Systems Committee (JISC) for two years, to develop a gateway to Internet resources in engineering as part of the UK Electronic Libraries Programme (eLib). It is based at Heriot-Watt University and is being developed in collaboration with six other universities in the UK (Moffatt 1996). *EEVL* is similar in concept to the subject services *OMNI* and *SOSIG*. The classification scheme adopted by *EEVL* is an in-house scheme which is loosely based on the Ei Classification. This approach was adopted after the investigation of a number of gateway services utilising conventional library classification schedules (UDC, Dewey, Ei) appeared to reveal that many parts of the subject trees remained empty. Rather than adopt an elaborate rigid classification which was originally developed for placing books on shelves as opposed to organising networked resources, the decision was taken to adopt a more fluid classification which could adapt as broad subject categories fill up.

Integration between classification scheme and controlled subject headings

All *EELS* resources are annotated, classified using the Ei classification and indexed with descriptors from the Ei Thesaurus. The Ei Thesaurus contains in its 2nd ed. 1995 over 8,300 descriptors and a total of 17,400 entry points including 9,000 non-preferred terms. It is a hierarchical thesaurus, which in addition provides the relationship between the descriptors and classification codes from the Ei classification. This allows *EELS* to offer sophisticated navigational and search features integrating browsing and searching, classification and thesaurus. Ei has so far not allowed *EELS* to display only the scheme to end-users (that is without any resources connected to codes).

Linking to third party classification data

The Ei classification codes and thesaurus serve as indexing tools for the database Ei Compendex*Plus, the printed Engineering Index and other index products. The main reason why *EELS* chose the Ei classification and thesaurus was the plan to build an integrated engineering service between Compendex for printed resources with *EELS* offering the digital resources, presenting to the user one single search and browse access point and retrieving records from both services. This development is not finished yet, since it requires changes to other software packages, like SilverPlatter's ERL and WebSPIRS.

Copyright

The Copyright for the classification system and the thesaurus belongs to Engineering Information Inc., Hoboken, N.J., USA.

Literature on Ei:

EELS: <URL:http://www.ub2.lu.se/eel/>

EEVL: <URL:http://eevl.icbl.hw.ac.uk/>

Ei home page: <URL:http://www.ei.org/>

Moffat (1996)

2.5.4.2. Mathematics Subject Classification

Usage

The 1991 Mathematics Subject Classification (MSC 1991) was compiled by the Editorial Offices of both *Mathematical Reviews* and *Zentralblatt für Mathematik / Mathematics Abstracts* which are the two main review journals in mathematics.

The American Mathematical Society (AMS) offers a Web page of *Materials Organized by Mathematical Subject Classification*. It lists main sections of the 1991 Mathematics Subject Classification with links to electronic journals, pre-prints, Web sites and pages, databases and other pertinent material in the corresponding fields: <URL:http://www.ams.org/mathweb/mi-mathbyclass.html>.

Availability

The scheme is available on the Internet: <URL:http://www.zblmath.fiz-karlsruhe.de/class/index.html>

Users are allowed to install a local version on their own machine: a tarred and compressed version is available at: <URL:http://www.ma.hw.ac.uk/~chris/mr-html.tar.Z>

Users of the system are asked to inform the authors so that they can inform of any updates.

Development effort

The editors of *Mathematical Reviews* and *Zentralblatt für Mathematik* have initiated the process of revising the 1991 Mathematics Subject Classification. They plan to have this revision completed by the end of 1998 so that it can begin to be used in *Current Mathematical Publications* in mid-1999, and in *Mathematical Reviews* and *Zentralblatt für Mathematik* from 2000.

2.5.4.3. ACM Computing Classification System (CCS)

The ACM (Association for Computing Machinery) <URL:http://www.acm.org/> Computing Classification System has become a standard for identifying and categorising computing literature, as well as areas of computing interest and/or expertise. The current taxonomy for categorising the computing literature saw its first release in 1982. Until recently, CCS was named the Computing Reviews Classification System (CRCS); it was renamed in recognition of its general use as a standard for classifying the computing literature. The 1991 Classification System is a cumulative revision of the 1982 version of the Computing Reviews Classification System. The 1982 Classification System had in turn superseded the previous CR classification introduced in 1964.

The Classification has two main parts: a numbered tree containing unnumbered subject descriptors, and a General Terms list. The unnumbered subject descriptors are essentially fourth level nodes.

Usage

It is Used in ACM's online databases and in CD-ROM files.

An Internet service using CCS is *Ariadne*, developed by the Medoc-project in Germany. <URL:http://ariadne.inf.fu-berlin.de:8000/>.

Availability

A complete copy of CCS is available in the January 1996 editions of *Computing Reviews* and ACM's *Guide to Computing Literature*. The CCS is also accessible on the WWW in both 1964 and 1991 versions: <URL:http://www.acm.org/class/>

Development effort

The ACM Publications Board is now updating the Computing Classification System (CCS).

Copyright

The copyright of the ACM Computing Classification System belongs to the Association for Computing Machinery (1995).

Literature on ACM CCS:

ACM home page: <URL:http://www.acm.org/>

3. Review of attempts to apply classification in automated services

3.1. Background

As classification is a time consuming and expensive process it is obvious that investigations into the use of automated solutions are worthwhile. At the same time, classification is an activity where a significant level of human expertise, abstract thinking and understanding is needed and this is not easy to substitute by artificial intelligence or expert systems.

There are no known examples of traditional library classification being overtaken completely by computer software.

However, knowledge structuring in the Internet has to cope with far larger numbers of documents, exponential growth rates and a high risk of changes occurring in documents that already exist.

This is the background for the development of a growing number of research projects and experimental systems, trying to support knowledge structuring activities on the Net. Most of these projects use methods of derived indexing, i.e. they extract information from the documents and use it for the structuring tasks.

Very few seem to make use of traditional library classification systems with universal or subject specific schemes constructed *a priori* over many years by co-operative organisations, independently from the contents of documents which actually exist in particular collections. This method is called assigned indexing, to devise an indexing language and assign the appropriate concepts or notations to each document.

Among the derived indexing methods and projects are:

- systems for text classification in USENET newsgroups or for video and image classification and other tools assisting in information structuring according to ad hoc or personal schemes (bookmark organisation, gathering and structuring collections)
- different clustering methods, based on statistical co-occurrence of words, citation links and co-citation structures or other similarity measures (The project HyPursuit at MIT Comp. Science Lab, <URL:http://www.psrg.lcs.mit.edu:80/Projects/CRS/HyPursuit> seems to foresee the possibility of using the LCC classification, amongst other vocabularies, as one kernel around which to cluster documents)
- linguistic methods using semantic clustering or (self-organising) concept maps
- ontology-based information capturing from the Net
- agents and learning systems (e.g. neural network methods) assisting browsing and searching (the largest group)
- construction of a topology of links between related resources for distributed scaleable searching

Many of those methods and tools could be used to improve quality controlled subject gateways although are not necessarily useful for building and structuring them from scratch.

Gerry McKiernan, Iowa State University, offers a comprehensive collection of pointers to such projects and systems, including short descriptions, citations and addresses following a study in the Summer of 1996 (McKiernan 1997). The clearinghouse is called *Project Aristotle(sm) - Automated Categorization of Web Resources* <URL:http://www.public.iastate.edu/~CYBERSTACKS/Aristotle.htm>.

In the following account, keeping to the focus of this report, only examples of methods and projects which stick closely to automatic routines connected with the usage of established library classification systems in Internet services are mentioned.

3.2. Automatic classification of WAIS databases using UDC

The Nordic WAIS/WWW Project at Lund University Library <URL:http://www.ub2.lu.se/W4.html> was the first project which tried to apply simple methods of automatic classification in order to improve the discovery and retrieval of Internet resources (from Summer 1993 to Summer 1994). It focused on a rather coherent single type of resource; about 700 WAIS databases, where it was possible to

automatically extract more or less detailed descriptions of their content from the Internet (Ardö, *et al*. 1994).

The part of the project that dealt with "Automatic indexing and classification of WAIS databases" <UTL:http://www.ub2.lu.se/autoclass.html> consisted of three elements:

a) Automatic detection of new WAIS databases
b) Automatic classification according to UDC (the English medium edition)
c) Construction of a WAIS subject tree based on the top levels of UDC.

WWW, WAIS and gopher front-ends provided access to the service from all three protocols.

An enhanced WWW/WAIS gateway allowed searching multiple WAIS databases in the same subject area simultaneously from HTML forms and a primitive relevance feedback <URL:http://www.ub2.lu.se/auto_new/UDC.html>.

Since early 1996, the service has not been updated. This is due to the fact that all directory-of-servers listing WAIS databases ceased their operation and WAIS databases today are "hidden" behind WWW pages, which makes it very difficult to discover them and collect any information about them.

The still usable automatic classification module of the software takes information from the keyword field and free-text parts of the selected database descriptions to construct a list of keywords for every WAIS database. This list is augmented with subject terms for the same database collected from a number of different sites on the net.

A list of suggested classifications is constructed by comparing the words from the collected descriptive vocabulary with UDC's vocabulary (although only a very limited part of it during the project).

When a match is found, the corresponding UDC classification (in the beginning restricted to the top level) is added to the list of suggested classifications with a weighting depending on which source the matching word originates from. Keywords in the subject field, for instance, have higher weights than ordinary words in the description field.

From the list of suggested classifications the final classifications are decided using a heuristic method based on the accumulated weights for each proposed classification and on the number of proposed classifications. Adjusting the weightings is essential in order to improve the quality of the classification process.

An evaluation of the outcome showed surprisingly few mis-classifications. In spite of efforts to use classifcation information from other sources, the result depended primarily on the quality of the original database descriptions provided by the authors/publishers.

The whole process is kept as automatic as possible. The methodology is not tied to UDC but could be used with other classification schemes, like the Library of Congress Classification, to produce different views of the resources.

3.3. Automatic classification of WWW resources in a robot-generated index using computer linguistic methods

Project *GERHARD* (German Harvest Automated Retrieval and Directory) is run by Oldenburg University Library: <URL:http://gerhard.bis.uni-oldenburg.de/>

GERHARD intends to create a robot-generated index of WWW resources in Germany and to automatically build a browsing structure by subject. It is being run by Oldenburg University Library, and runs for one year (until the end of 1997; a prototype is expected by May 1997) and supported by the Deutsche Forschungsgemeinschaft (DFG).

GERHARD uses a similar, but linguistically more advanced method than the Nordic WAIS-WWW Project did and applies it to a much larger and more heterogeneous set of documents.

Computer linguistic methods are used (as developed by the project partner ISIV, the Institute of Computer Linguistics, Osnabrück University) to match the natural language content of the indexed documents to a library classification system.

The enlarged and multilingual version of UDC used at the ETH library Zürich is chosen as the most suitable classification system for the task since it is optimised for computer use and display (Loth 1996).

The automatic classification process consists of two parts: the computer linguistic analysis and the matching to the specially prepared UDC scheme. The natural language content of HTML pages is segmented into suitable entities, words and phrases, and compared with a dictionary created from the UDC system. The comparison results in a set of UDC notations for every document which are ranked and weighted statistically according to frequencies and to the structure of the document.

The relevant parts of the documents contents are indexed, together with the resulting classification notations, into a database open to direct searching. In addition a UDC subject tree for all documents is dynamically generated and provided as a browsing structure.

This offers the possibility to integrate the index and the browsing structure to allow the user to jump from individual hits in the search results to the wealth of related documents in the proper sections of the classification system (the same feature is offered in the *EELS* service).

3.4. Automatic classification and indexing using DDC (Project Scorpion, OCLC)

The most important project in the area of automatic classification is OCLC's research project Scorpion <URL:http://purl.oclc.org/scorpion/>.

"Scorpion is a research project at OCLC exploring the indexing and cataloging of electronic resources. Since subject information is the key to advanced retrieval, browsing, and clustering, the primary focus of Scorpion is the building of tools for automatic subject recognition based on well known schemes like the Dewey Decimal System (DDC)" (Shafer 1996).

In Scorpion, the document requiring indexing is treated as a query against a DDC knowledge base consisting of the vocabulary, notations, structure and relations of the classification system. The input terms for the query varies according to the type of documents to be classified, from title and subject headings only to including large parts of the documents' vocabulary. The results of the search, using ranked retrieval, will then become the subjects of the document and can then for instance be used by a human cataloguer for classification support. The knowledge base for DDC is the Electronic Support System (ESS) used by the *Electronic Dewey for Windows* CD-ROM product from OCLC Forest Press (cf. 2.1 The Dewey Decimal System). The intention is to increasingly exploit linkages between Dewey classes and LC Subject Headings as well as other subject schemes gathered from different databases and conversion programs.

So far, the project has explored the use of DDC as a concept definition source. DDC was deemed very suitable for the task because it demonstrated a high degree of class integrity: meaning that subject definitions are unambiguous and well-defined. DDC would therefore be a suitable system for automatic classification. (Thompson. *et al*. 1997)

A similar study of the characteristics of LC Classification could build upon the older experiments by Larson (1992). He tried to use partial-match retrieval techniques between various elements of new records and a database of classification clusters generated from titles and subject headings from previously classified MARC records. Between 46% and 86% of the books were classified correctly. The outcome of the experiment was to recommend a semi-automatic method of classification for books.

3.5. Improvement of classification by Neural Network techniques

KBS_media Lab, Lund: <URL:http://delphi.kstr.lth.se/kbs/projects/kbscross.html>

There are a couple of projects trying to apply neural network technology in order to improve the learning behaviour of software supporting navigation, browsing and organising systems. The project KBS-CROSS is attempting to develop automatic cross-referencing between classification systems with neural networks.

The goal is to produce a computerised tool that cross-references the LC Classification system with the UDC within the domain of building and architecture.

The practical application will be to allow Lund University Library's cataloguers and indexers to index documents faster and more accurately by suggesting UDC notations based on the LCC/LCSH data already associated with the document. Conversely, literature searchers will be offered use of a LCC/LCSH system for searches in literature databases indexed with UDC. An OPAC or another search system would be enabled by this tool to properly handle different knowledge representations.

Data from the Swedish union catalogue LIBRIS is used together with neural network and text processing tools. The resulting code and data is deployed in a demonstration tool in a Dynamic Knowledge Net. The tool's performance is being continuously evaluated by indexing specialists.

3.6. Automatic conversions between classification schemes

The possibility to automatically convert from existing classifications of documents (OPAC or database records, documents in Internet services, etc.) into another scheme used in a quality controlled subject gateway could become a potentially valuable support for the classification task. This method is occasionally used in co-operative cataloguing projects and union catalogues, sometimes even in individual OPACs as soon as cataloguing records using a different classification scheme are imported or exchanged.

Examples:

Pica is automatically converting Library of Congress classification into the Nederlandse Basisclassificatie (BC) notations when LC records are loaded into the Pica OPAC. Similar software is used as a concordance from UDC to BC.

For the Swedish national classification system SAB, mostly used by arts, humanities and social science libraries in the research libraries union catalogue LIBRIS, there exists a printed conversion table from Dewey to SAB. However, it is not very up to date (DDC ed. 19. and SAB 6. ed., 1987).

If there are no "official" conversion tables available, an improvement of the classification task could still be reached by extracting linkages between different classification systems (possible in LIBRIS) or between indexing terms and classification for the same object from existing databases (LC catalogue; *EELS* etc.) and use it as conversion algorithm.

Ingetraut Dahlberg’s idea at a panel during the Thirty-sixth Allerton Institute (Wyly 1995, p. 77) to offer the Internet searcher a black-box, containing a range of concordances between classification schemes, exploiting classification data from different systems for improved search support, would be the ultimate step as far as conversion is concerned.

She had already proposed in 1982 an outline of a so called "Information Coding Classification" as a first step to an effort for a new general system, as a global switching mechanism between all classification systems and the databases that make use of them (Dahlberg, 1995, p.31-34)

3.7. Summary

Automatic classification processes are a necessity if large robot-generated services are to offer a good browsing structure for their documents or advanced filtering techniques as well as proper query expansion tools to improve the search process.

Even if we restrict ourselves to the use of *a priori* classification of documents with traditional library schemes there are a couple of promising approaches. However, large-scale reliable systems are not to be expected in the immediate future.

Quality controlled subject gateways with limited size could use methods of shared classification and scheme conversions. The work of the human cataloguers and classifiers could be improved by many of the above mentioned techniques and tools as supportive, semi-automatic systems. Clustering and selection methods, agents and self-configuring maps are the most promising tools as are the possible outcomes of project Scorpion.

4. Conclusion

The Thirty-sixth Allerton Institute, held in October 1994, was a starting point for the discussion of the use of classification systems in information networks. Most of the closing remarks by Marcia Bates and Sarah Thomas pointing to important working directions are still relevant to the issues covered in this report (Cochrane 1995):

"1. Exploit technology

a) for adding class numbers to materials in digital form

b) for linking subject access systems like LCSH and DDC.

c) for providing navigation and retrieval tools based on outlines of knowledge within classification schedules.

2. Extend the use of library classification to Internet resources …

4. Share development strategies among and between various classification systems and thesauri, creating the ability to link with one another including multilingual and specialized systems …

6. Build bridges from the past (e.g. , library collections classified by DDC, LCC, etc.) to the future (e.g., digitized full text collections) …

12. Organize the classification schemes differently for the end-user than for the classifier and provide more than one scheme for users to browse and navigate before and after retrieval".

In the meantime, a variety of classification schemes are being used to bring systematic order to discovery oriented Internet services. The major universal schemes like DDC and LCC are mainly used by services run by the library community, while UDC is used primarily in Europe for subject specific services or for a general information gateway like *NISS.*

For services in several countries, like the Netherlands or Sweden, national general schemes are used. Subject specific international schemes are the dominating choice among subject-based information gateways. However, many services develop browsing structures on their own, similar to traditional classification systems or have developed extensive local adaptations of existing schemes.

TABLE 4.1: Summary of reviewed classification schemes

	BC	**CCS**	**DDC**	**Ei**	**Icon.**	**LCC**	**MSC**	**NLM**	**SAB**	**UDC**
Number of Internet services using system	1	1	17	2	0	5	0	1	4	5
Multilingual capability	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Widely translated.	No	No	Yes	No	No	No	No	Yes	No	Yes
Integration with LCSH	No	No	Yes	No	No	Yes	No	No	No	No
Integration with other systems	GTT	No	LCC	Ei the-saurus	No	DDC	No	MeSH	No	No
Digital Availability	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	No	Yes
Copyright	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Extensibility	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Subject-specific	No	Yes	No	Yes	Yes	No	Yes	Yes	No	No

Table 4.1 shows some of the features identified in section 2 for all the reviewed classification schemes. The most used scheme in Internet services is DDC, which reflects its use in traditional and other online services. All the schemes have a multilingual capability to the extent that they use Arabic numerals, sometimes with added letters from the Latin alphabet. The real constraint on their use, however, is the availability of suitable translations and only UDC and DDC have been translated to any significant degree. Both LCC and DDC are integrated to some extent with LCSH and other schemes are integrated with relevant subject schemes, like NLM to MeSH and Ei to the Ei thesaurus. Most of the schemes are

available in some digital form, although the exact way this is done varies between schemes. All the classification schemes are extensible, although not always in a completely logical manner, e.g. new subjects are fitted into any remaining gaps in SAB.

The use of a classification scheme in an subject-based Internet service would be extremely useful. It offers the following advantages:

- It brings together small collections of similar resources
- The use of a systematic well-supported hierarchical structure supports the browsing of these collections
- It gives a context for search terms and allows filtering and high precision searches
- If the same classification scheme is used, more than one database can be searched with the same approach

The main criteria for the choice of classification system would normally be the scope of the service: its subject, language and geographic coverage and its user population.

In some situations the solution is quite obvious: for documents from all areas of knowledge, published throughout the world and in many languages and to be offered to an international multi-disciplinary community of users, a universal scheme can be selected, at least as a basic solution. DDC and UDC have a good multilingual capability due to the fact that they are entirely numerical and their schedules have been widely translated. If the collection however focuses on a rather limited subject area or discipline and there is a suitable international subject-specific scheme available, it should be used.

Problems will occur for services covering subjects where there are several different schemes (e.g. the earth sciences), although the use of concordances may help. There will also be problems when there is no comprehensive scheme available for a service covering a particular geographic area or subject scope (e.g. the European social sciences in *SOSIG*).

Perceived shortcomings in classification schemes are sometimes countered by adaptations and amendments to a scheme. For example: *EEVL*'s variant of Ei, *NISS* and *SOSIG*'s use of UDC, etc. Adaptations can arise from the use of classification schemes in this different electronic environment. One is not preparing a shelf arrangement of physical objects, but a digital, virtual display in an online system where the classification scheme itself is used as a browsing aid.

Another reason for adapting classification schemes is the potential, when using the exact version of a library classification system, that some parts of the scheme could remain completely empty while other parts of the scheme are overcrowded. This is due to the possibility that the subjects in existing digital documents might widely differ from those found in printed collections, or that the sizes of printed and digital collections in this subject area might also be different (cf. 2.5.4.1. Ei classification).

In spite of these good reasons to locally adapt schemes, changes to a scheme will hamper interoperability and co-operation.

Interoperability between subject services could be accomplished by an hybrid usage of universal and subject-specific schemes. Universal schemes could 'glue' different subject systems together and provide a coherent structuring principle at a top-entry level to subject specific services. Then, when moving into the subject services themselves, a subject-specific scheme could be used.

With regard to subject-specific classification schemes, it is advisable that only well-established schemes should be used. Whenever feasible, especially in small services, it might help if a classification from one of the universal schemes could be added. Conversion programs between classification schemes could help accomplish interoperability as well.

Home-grown schemes on the Web are not normally specifically designed to classify academic resources (for the research community) but aim to categorise a wider breadth of form and content: e.g. entertainment, commercial information, government information, etc. UDC, Dewey and LCC and subject-specific schemes, on the contrary, have been developed as schemes to classify the whole of knowledge and are especially useful for classifying academic resources, although as DDC shows, they can embrace popular types of content too. For an academic subject service, home-grown schemes should, therefore, not be developed.

Scheme conversion programs and methods of shared classification are considered very useful especially for subject specific services. Different methods of derived indexing recently developed, clustering and

selection technologies, agents and concept maps, and similar techniques of automatic classification are soon expected to offer good improvement in services of limited size.

Appendix 1: E-mail Questionnaire sent to Internet Services using UDC

Responses were received from:

SOSIG: Emma Worsfold/Debra Hiom
NISS: Oren Stone
BUBL: Dennis Nicholson
OMNI: Sue Welsh

The questionnaire:

Dear ,

We would like to include your service in a review being carried out by the EU-funded DESIRE Project, and would be very grateful if you could answer a few questions for us.

DESIRE is surveying the use of classification schemes in Internet Information Services, and I am aware that you are currently using the UDC (Universal Decimal Classification). Please could you answer the following questions and return them to me? Many require only a yes/no answer so I hope it will not take up too much of your time. Many thanks!

1) Have you had to pay for the right to use the scheme, or to 'publish' it in displays?

2) Do you have a version of your classification scheme in electronic form? (Either a complete version , or an abridged version - are you able to give a URL?)

3) Have ever made use of the fact that the scheme is available in different languages?

4) Does your service integrate the classification scheme with other systems such as a thesaurus or controlled subject headings?

5) At what level of detail do you classify?

6) What do you consider to be the strengths of the scheme in relation to its use in your service? (i.e. Why did you chose this scheme as opposed to others?)

7) What do you consider to be the weaknesses of the scheme in relation to its use in your service?

8) Have you found any strengths/weaknesses of the scheme in particular subject areas? (i.e. is it better for some subject areas than for others?)

9) Are there any other points you would like to make about your use of a classification scheme?

Many thanks for your time.

PART IV: Bibliography and References

Ardö, A., Falcoz, F., Koch, T., Nielsen, M., Sandfaer, M., 1994, Improving resource discovery and retrieval on the Internet. The Nordic WAIS/World Wide Web Project - Summary report. *NORDINFO Nytt* , 4, 13-28. <URL:http://www.ub2.lu.se/W4/summary.html>

Art, Design, Architecture and Media Information Gateway (ADAM), 1996, *1996 Survey of user information needs and search methods results*. <URL:http://adam.ac.uk/adam/reports/survey/>.

Brandhorst, H. and Van Huisstede, P., 1992, Iconclass in de computer - De classificatie van beeldmateriaal in een geautomatiseerde omgeving. *Open*, 24 (9), 294-298.

Buchanan, B., 1979, *Theory of library classification*. London: Bingley.

Buxton, A., 1990, Computer searching of UDC numbers. *Journal of Documentation*, 46 (3), 193-217.

Chan, L.M., 1986, Library of Congress Classification as an online retrieval tool: potentials and limitations. *Information Technology and Libraries*, 5 (3), 181-192.

Chan, L.M., 1995, *Library of Congress Subject Headings: principles and application*, 3rd ed. Littleton, Colo.: Libraries Unlimited.

Cochrane, P.A., 1995, New roles for classification in libraries and information networks. In: New roles for classification in libraries and information networks: reports from the Thirty-sixth Allerton Institute. *Cataloging and Classification Quarterly*, 21 (2), 3-4.

Comaromi, J.P., *et al*., 1989, *Dewey decimal classification and relative index: devised by Melvil Dewey*, 20th ed. 4 vv. Albany, NY.: Forest Press.

Comaromi, J.P., *et al*., 1990, *Abridged Dewey Decimal Classification and relative index*, 12th ed. Albany, N.Y.: Forest Press, 1981.

Dahlberg, I., 1995, The future of classification in libraries and networks: a theoretical point of view. In: New roles for classification in libraries and information networks: reports from the Thirty-sixth Allerton Institute. *Cataloging and Classification Quarterly*, 21 (2), 23-36.

Dodd, D.G., 1996, Grass-roots cataloging and classification: food for thought from World Wide Web subject-oriented hierarchical lists. *Library Resources & Technical Services*, 40 (3), 275-286.

Facharbeitsgruppe Sacherschliessung, 1995, *Basisklassifikation für den Bibliotheksverbund Niedersachsen/Sachsen-Alhalt/Thüringen*. Zweite überarbeitete Ausgabe, überarbeitet durch die Facharbeitsgruppe Sacherschliessung.

Foskett, A.C., 1973, *The Universal Decimal Classification: the history, present status and future prospects of a large general classification scheme*. London: Bingley.

Grund, A., 1993, Iconclass. On subject analysis of iconographic representations of works of art. *International classification*, 20-29.

Guenther, R.S., 1992, The development and implementation of the USMARC Format for classification data. *Information Technology and Libraries*, 11 (2), 120-131.

Guenther, R.S., 1996, Automating the Library of Congress Classification Scheme: implementation of the USMARC Format for Classification Data. *Cataloging & Classification Quarterly*, 21 (3/4), 177-203.

INFOMINE, 1996, *About INFOMINE* <URL:http://lib-www.ucr.edu/infomine/intro.html>

Langridge, D., 1973, *Approach to classification: for students of librarianship*. London: Bingley.

Langridge, D., 1991, Classifying knowledge. In: Meadows, A.J., (ed.), *Knowledge and communication: essays on the information chain*. London: Library Association Publishing, 1-18.

Larson, R.R.,1992, Experiments in automatic Library of Congress Classification. *Journal of the American Society for Information Science*, 43 (2), 130—148.

Loth, K., 1996, Wissensorganisation durch ein neues Notationssystem - eine konstruktive Kritik der UDK. *ABI-Technik* 16, 1.

Marcella, R. and Newton, R., 1994, *A new manual of classification*, Aldershot: Gower.

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.

McIlwaine, I.C., 1991, UDC as a standard for subject control. In: McIlwaine, I.C., (ed.), *Standards for the Exchange of Bibliographic Information: papers presented at at course held at University College London, 3-18 August 1990*. London: Library Association.

McIlwaine, I.C., 1993, *Guide to the use of the UDC*. FID Occasional Paper 5. The Hague: International Federation for Information and Documentation (FID).

McIlwaine, I.C., 1995a, UDC Centenary: the present state and future prospects. *Knowledge Organisation*, 22 (2), 64-69.

McIlwaine, I.C., 1995b, Preparing traditional classification for the future: Universal Decimal Classification. In: New roles for classification in libraries and information networks: reports from the Thirty-sixth Allerton Institute. *Cataloging and Classification Quarterly*, 21 (2), 49-58.

McIlwaine, I.C. and Buxton, A., 1995, *Guide to the use of UDC: an introductory guide to the use and application of the Universal Decimal Classification*, rev. ed. FID Occasional Paper 5. The Hague: International Federation for Information and Documentation (FID).

McKiernan, G., 1996, *Beyond bookmarks: schemes for organizing the Web*. <URL:http://www.iastate.edu/~CYBERSTACKS/CTW.htm>

McKiernan, G., 1997, Hand-made in Iowa: organizing the Web along the Lincoln Highway. *D-Lib Magazine*, February. <URL:http://www.dlib.org/dlib/february97/02mckiernan.html>

Mitchell, S., *Library of Congress Subject Headings as Subject Terminology in a Virtual Library: The INFOMINE Example* <URL:http://lib-www.ucr.edu/pubs/postlcsh.html>

Mitchell, J., 1995, DDC 21 and beyond: the Dewey Decimal Classification prepares for the future. In: New roles for classification in libraries and information networks: reports from the Thirty-sixth Allerton Institute. *Cataloging and Classification Quarterly*, 21 (2), 37-48.

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

Moffat, M., 1996, An *EEVL* solution to engineering information on the Internet. *Aslib Proceedings* 48 (6), 247-258 <URL:http://eevl.icbl.hw.ac.uk/paper1.html>

Nederlandse Basisclassificatie, 1992, *Nederlandse Basisclassificatie* [Stichting voor Bibliotheekautomatisering Pica en het Samenwerkingsverband UKB]. Tweede herziene uitgave. Leiden: Pica ; Amsterdam: UKB.

New, G. and Trotter, R., 1996, Revising the life sciences for Dewey 21. *Catalogue & Index*, (121), 1-6.

NISS, 1996, *How to fill in your NISS information gateway templates: the big user guide*: 8. Library of Congress Subject Headings (optional). <URL:http://www.niss.ac.uk/resource-description/bigguide.html#8>

Odlyzko, A.M., 1995, Tragic loss or good riddance? The impending demise of traditional scholarly journals. *International Journal of Human-Computer Studies*, 42, 71-122.

Oehler, A., 1996, *Browsingsdienste im Internet*. <URL:http://fub46.zedat.fu-berlin.de:8080/~angela/bond/browsing.htm>

Ranganathan, S.R., 1965, *The Colon Classification*. Rutgers series on systems for the intellectual organization of information, 4. New Brunswick, NJ.: Rutgers State University, Graduate School of Library Service.

Rowley, J.E., 1987, *Organising knowledge: an introduction to information retrieval*. Aldershot: Gower.

Schatz, B., Mischo, W.H., Cole, T.W., Hardin, J.B., Bishop, A.P. and Chen, H., 1996, Federating diverse collections of scientific literature. *IEEE Computer*, 29 (5), 28-36.

Seattle Pacific University Library, 1996, *Introduction to Library of Congress Classification* <URL:http://www.spu.edu:80/depts/library/second/lcc.html>

Shafer, K. (1996). *A brief introduction to Scorpion*. Dublin, Ohio: OCLC. <URL:http://orc.rsch.oclc.org:6109/bintro.html>

Steinberg, S.G., 1996. Seek and ye shall find (maybe). *Wired* [U.S], 4.05, May. <URL:http://www.hotwired.com/wired/4.05/features/indexweb.html>

Svenonius, E., 1983, Use of classification in online retrieval. *Library Resources & Technical Services*, 27 (1), 76-80.

Thompson, R., Shafer, K. and Vizine-Goetz, D., 1997, *Evaluating Dewey concepts as a knowledge base for automatic subject assignment*. Dublin, Ohio: OCLC. <URL:http://orc.rsch.oclc.org:6109/eval_dc.html>

U.S. National Library of Medicine, 1996, *NLM Classification* (factsheet). <URL:http://www.nlm.nih.gov/publications/factsheets/nlm_classification.html>

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. <URL: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 organisation and change: proceedings of the 4th international ISKO conference, 15-18 July 1996, Washington D.C*. Frankfurt/M: INDEKS Verlag. <URL:http://orc.rsch.oclc.org:6109/dvgisko.htm>

Vizine-Goetz, D. and Markey, K., 1989, Characteristics of subject heading records in the machine-readable library of Library of Congress Subject Headings. *Information Technology and Libraries*, 8 (2), 203-209.

Wallace, J. and Burden, P., 1995, Toward a classification-based approach to resource discovery on the Web. Wolverhampton: University of Wolverhampton, School of Computing and Information Technology. <URL:http://www.scit.wlv.ac.uk/wwlib/position.html>

Wyly, B., 1995, What lies ahead for classification in information networks? Report of a Panel Discussion. In: New roles for classification in libraries and information networks: reports from the Thirty-sixth Allerton Institute. *Cataloging and Classification Quarterly*, 21 (2), 75-82.

Wynar, B.S., 1992, *Introduction to Cataloguing and Classification*, 8th ed. Englewood, Colo.: Libraries Unlimited

Also:

Private e-mails:

From Rebecca S. Guenther, Debra Hiom, Karen Oberst, Matt T. Rosenberg, and Richard Sapon-White. to Anna Brümmer (Dec 1996 -Jan 1997)

From Winnie Kao <Winnie_Kao@occshost.nlm.nih.gov> to Marianne Peereboom (KB)