Crossing the border
Subject search across library catalogues - attempting to match subject descriptions by a quantitative method

Bjarne Andersson and Knut Hegna*

Originally published in VINE, no. 114 (1999), p.56–66.
This version slightly modified (layout)
when transformed to LaTeX October 2003.

Abstract: This article is a preliminary attempt to discuss and construct an experimental library user interface, focusing on connecting different libraries in such a way that each library can search other library collections using its own classification codes. The second focus of this article is viable method to locate specific subject clusters in a library catalogue. The classification codes need not be uniform, and the user interface allows for natural language searching. The technique used to obtain the above mentioned is: construction of a concordance table between different library catalogues' subject codes on the bibliographic level, enabling the system to determine relative links between generic different subject codes. The concordance is established on the basis of shared titles.

Contents

Bjarne Andersson - Ass. prof. and subject specialist, Roskilde University Library, Roskilde. *Knut Hegna* - Senior academic librarian, University of Oslo, Informatics library.
His work on this article has partly been funded by the Ministry of Culture in Norway and Astrup-Hoels minnefond.

1 Introduction

The work presented in this article was initiated during a joint visiting period, where both authors[1] worked at the research and development departement at the Technical Knowledge Center and Library of Denmark[2].

We were both working on improvements of subject based user interfaces for searching. Not that this kind of improvement should take much of our time according to the reviews libraries was getting on this specific topic, but it turned out to be a lot more complicated than we anticipated.

Our experience is that most libraries have a rather non-sophisticated approach to subject data, as seen from most OPAC user interfaces. Some systems hide this bibliographic structure from public use, leaving the structure as a magic tool of the professional staff.

Other systems fail to mention that their subject codes are quite unique in the library society as such, and therefore almost non-usable for users not accustomed to the local rite. You seldom see the verbal code description used as an integrated tool in catalogue searching.

As we plunged into the work we realised that most of the (important) problems weren't even on the agenda for the discussions in the library society in our part of the world. This will hopefully become apparent in the following paragraphs, where we will try to elaborate on this statement.

1.1 Subject description in libraries

Our critique could very easily become a tiresome rattling off of trifles, and we will be the first to apologize for that, but to strenghten our later argument you have to bear with us for a moment.

- Whenever we try to discuss "subject classification" we get the notion that this form of description is a kind of art work and therefore not subject to quantitative arguments.
- Most libraries are pleased with their own subject classification method, and wouldn't dream of changing it. This conservatism originates in a tendency to avoid obvious laborious work with little local benefit, i.e. search-facilitating-features. This fact seems to originate in a tendency to acquiesce to already made decisions - once you have choosen a classification code, and started to make hand-outs and user guides

[1] Unfortunately both our works is published in Danish / Norwegian. If this fact isn't discouraging see Bjarne Andersson: Arbejdsrapport: Virtuelle specialbiblioteker. En Webbaseret model for samtidig emnesøgning i forskellige bibliografiske databaser. Roskilde Universitetsbibliotek, 1998. URL http://www.rub.ruc.dk/~bas/virt/arap_vir.html - Knut Hegna: Rapport fra studie- og arbeidsopphold ved Danmarks Tekniske Videncenter og bibliotek (DTV). Nordinfo-nytt, nr. 1/2, 1998, s.33-50. URL http://www.ifi.uio.no/~knuthe/dok/dtv.html

[2] DTV - URL http://www.dtv.dk/

incorporating these codes, why should you change it? *Though this be madness, yet there's method in't.*

- The medium and large libraries are prone to scientific "slumps" where the intensity and enthusiasm of the classification personnel dries up, thus creating areas of obsolete or inadequate subject codes. This could also be an effect of the fads and foibles in the scientific field. When nobody except the staff uses subject code search, why bother about it, why put any effort into it ?

- Most subject code systems are typical of a given period. UDC[3] for example was a postwar endeavour to unite the split in the scientific fields (between east and west, between natural and social sciences) and had never the less a fair amount of scientific peculiarities[4], with an additional possibility of attendant controversies.

- Subject classification by natural language words (either by a thesaurus or by a more informal structure) is not a panacea for the muddle in the subject description area. The word 'Class' could very well cause a major misinterpetation in a system where all labour movement related topics were classified in a business management dominated scheme.

Hopefully this list should suffice for our point - subject classification needs a helping hand, especially in the case where several libraries try to make a web-based virtual (we know, that word again) catalogue. Subject classification has (at least) two distinct levels of meaning. One is "The simple level" denoting the specific library and its effort to make materials traceable for future use. The other is a "Meta level" where the subject classification is a kind of key to the overall resources of a specific body of libraries.

Both levels are imporant for the making of virtual catalogues, but our aim in this article is a discussion mainly of the first, although descriptions of larger sets of similar materials (like a package of electronic periodicals) will need a meta level pointer.

In addition to this point we acknowledge that accurate and consistent classification makes the library catalogue a scientific tool, but a discussion along this line falls outside our present scope.

1.2 Sharing data

When this is said, another important factor, and an almost tacit knowledge in the library society, has to be mentioned. One of the not so popular activities in the library sector is sharing data. This sounds odd, inasmuch as one of

[3]URL http://http://www.udcc.org/about.htm

[4]In the 1967 edition of UDC you had to look for womens lib in the folklore section on the same level as rules of etiquette and primitive people.

foremost priorities of especially the scientific libraries is providing access to data, information, etc. The providing is more often than not viewed as a one way process where the specific library (kind of) owns its local data, books etc. - even if it participates in a joint venture to facilitate this providing process.

This state of public privacy has specific consequences for the data handling in libraries. Most agree on the appropriateness of joint standards for traditional descriptive bibliographic information like author, title, imprints data, etc.

When we reach subject codes - agreement is swapped with local chauvinism and ditto practice. Everybody uses a kind of universal classification scheme (only adjusted a tiny bit to the local needs) and has great difficulties communicating with other libraries on this level. Or they use "natural language" in form of either a thesaurus or a loosely connected heap of subject words / subject headings.

The different practices might originate from differences in the collection building policy also. Varying classification depths for the same document might depend on the weight this particular subject has in the collection as a whole.

In a specific library you can always see your way through or around the subject classification problem. Either by instigating user courses or by extensive on-site-information. The local subject description pecularities were acceptable when you looked at the collection and its catalogue as parts of the same isolated island. Even though it might contain madness, the consistency could be perfect. But in a cooperative, joint, area-wide, or global library system - what is going to happen?

Is a standard communication protocols (like z39.50) the answer to this dilemma? Hardly.

Standard communication protocols are of great use where you need a fast and reliable data communicating process (e.g. between several libraries) capable of exchanging and merging bibliographical, holding, and circulating data. But this strength vanishes in thin air as soon as you start talking subject classification codes, that is unless all the participating libraries agree on applying the same classification system, preferably in the same manner. If libraries use different (local) subject codes, you are back to square one, getting no benefit from the standard communicating protocol.

2 Two libraries - one subject search facility

We decided therefore to investigate a user interface connecting our two libraries in such a way that each library could search both library collections using its own classification codes. Roskilde University Library (RUb) uses a local version of the universal UDC, while The Informatics Library (IFI) at

ISBN	CRCS classification	UDC classification
0-273-08665-0	I.2.1	681.3:159.9 681.3.01 61
0-273-08667-7	I.2.1 Ø.8.0	681.3:159.9 519.72
0-306-30801-0	I.2.6 I.2.9	159.955.6 681.3:159.9 061.3:159.955.6
0-306-44131-4	A.m	001:3 371.12:378 396
0-333-25749-9	C.1.0 B.3.0 B.6.0	681.32 681.3
0-385-26774-6	H.1.2 K.6.1 H.5.2 A.1.0	745 613 159.98
0-385-41993-7	K.4.0	681.3 301.151:007 301.152TELE

Table 1: Class codes for shared titles.

The University of Oslo (UiO) uses a local version of The ACM Computing Classification System (CRCS). RUb is in principle a "universal library" covering almost every scientific subfield and IFI is a hardcore computer science library.

2.1 Shared titles

To handle this mutual and on the same time locally based search facility we had to develop a table of concordance between classification codes from the respective catalogues.

We have seen several attempts to create concordance between classification systems, some with a very good graphical representation[5]. All these models are designed to join and evaluate the different classification structures at a specific library, where the overall classification practice is likely to be a "fixed entity" regardless of the applied classification system. This is unfortunately not the case in our situation. We have to make allowance for different classification practice co-existing with different classification systems.

We tried to overcome this obstacle by constructing our table of concordance by working only with the joint stock of monographs[6], and thereby comparing the classification codes of each item. This boils down to a purely quantitative method.

The shared titles were identified by comparing the ISBNs from the catalogue of the Informatics library (roughly 6000 ISBNs) to the ISBN-registry of the RUb catalogue (roughly 246000 ISBNs). This comparison resulted in the identification of about 1000 shared unique titles.

In table 1 you see some examples of shared titles (represented by ISBN) and the corresponding subject codes from both libraries[7].

[5]Fex. between DK5, DDC og FMB/UDK at the State and University Library, Aarhus, Denmark. A graphical representation is scheduled to be in use late 1999.

[6]A subset of a litle over 1000 records. This is not enough records to make a fully satisfactory table of concordance, but we have to stick to it for the time beeing.

[7]RUb's total monographs: 250000. IFI's total monographs: 13000.

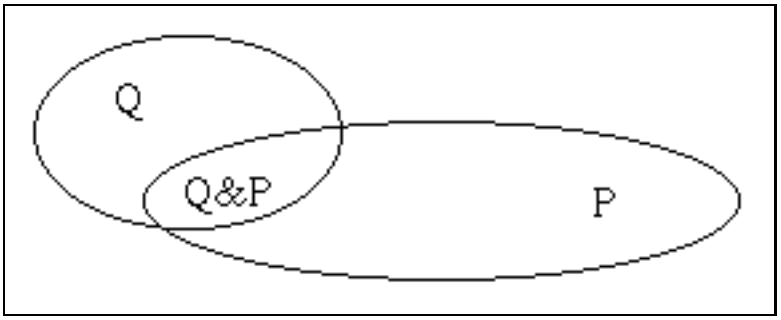

Figure 1: The relationship between two classification schemes

2.2 The concordance table

Most of the establishing of the concordance table revolved around the concept: relative frequences or relations. That is, relations between different subject codes from different libraries. Let us as an illustration look at the subject 'history of computer science'. This is in UDC mainly represented by P: a code called 681.3(091) while CRCS usually represents it by Q: code K.2.0 ('history of computing'). If we seek the relative relation between those two codes (by the joint stock of monographs) we have the boolean diagram in figure 1 where Q & P denotes bibliographic records having both the Q and the P code attached, whereas Q represent the records with only that code, and likewise for P[8]. If we try to determine the relative relation between Q and P it will be asymmetric, one for the relation Q $\rightarrow$ P and another for the relation P $\rightarrow$ Q. A first approximation of these relations could be:

Q $\rightarrow$ P to a degree of $\frac{\#(Q\&P)}{(\#(Q)+1)}$ and

P $\rightarrow$ Q to a degree of $\frac{\#(Q\&P)}{(\#(P)+1)}$

given that $\#(Q)$ is the number of bibliographic records described by code Q, $\#(P)$ is the number of bibliographic records described by code P, and $\#(Q\&P)$ is the number of bibliograpic records described by both the codes[9]. We add one to the denominator simply to decrease the importance of low-frequent codes.[10]

[8]It would be relevant at this point to stress the nescessity of making code-comparisons inside the local code system as well as between the different code systems. UDC code 681.3(091) relates to several other UDC codes with a strong relative frequence. However, it is not crucial for the illustrative purposes of this article.

[9]This is a preliminary function. It is used for instance by Troels Andreasen & Tommy Shomacker: *DabBib –a union catalogue applied for user friendly flexible querying*, paper presented at IFLA 1997 Congress.

[10]If we wanted to increase or decrease the importance of low-frequency codes we could

UDC	CRCS	Relation
681.3(091) →	K.2.0	0.79
	K.8.0	0.07
	K.4.0	0.07
	K.1.0	0.07
	I.2.0	0.07
	H.0	0.07
	D.3.2	0.07
	C.5.3	0.07
	C.0	0.07
	A.0	0.07

Table 2: Relations between a single UDC code and CRCS codes (P → Q)

Both libraries use multiple classification, i.e. several codes describing each title. Since citing order of codes has no significance, it is hard to establish the right relations between codes on that basis. The citing order is more or less random. In CRCS two codes might describe the joined subjects of the codes (History of programming languages = K.2.0 and D.3.2) or the separate treatment of two different subjects in the same book. There is no way to differentiate between the two practices. And where the UDC-system in special cases uses the colon notation to join subjects, the CRCS uses two separate codes, so we can't really establish the right relation between codes unless we accept that each code from one system might relate with higher or lower weight to several codes in the other system.

As most classification codes relate to several other codes, we made a computation of all the relevant relations. In tables 2 and 3 you will se the relations for code P – 681.3(091) – and code Q – K.2.0 – with the corresponding frequencies (all relations):

The next step was to assemble and count all occurencies of each code from both systems and compute the relative frequencies of mutual co-occurrences. In tables 2 and 3 you see a few lines from the raw data as generated from data on the table 1 form.

2.3 The user interface

Functionality

The first step in this interface[11] is searching for a subject in one of the the two libraries' classification systems (chosen by pull-down menu, see figure 2).

also use an exponential solution like the one presented in Reginal Ferber: *Automated indexing with Thesaurus Decscriptors: a Cooccurence Based Approach to Multilingual retrieval*, in Lecture notes in computer science, no.1324, p.128.

[11] Starts at URL http://www.rub.ruc.dk/~bas/eold/search2txt.

CRCS	UDC	Relation
K.2.0 →	681.3(091)	0.35
	62(091)	0.26
	681.1	0.19
	338.45:621.38	0.16
	62:3	0.13
	681.32	0.10
	621.38:338.45	0.10
	681.3:159.9	0.06
	621.382	0.06
	001.89	0.06

Table 3: Relations between a single CRCS code and UDC codes (Q → P)

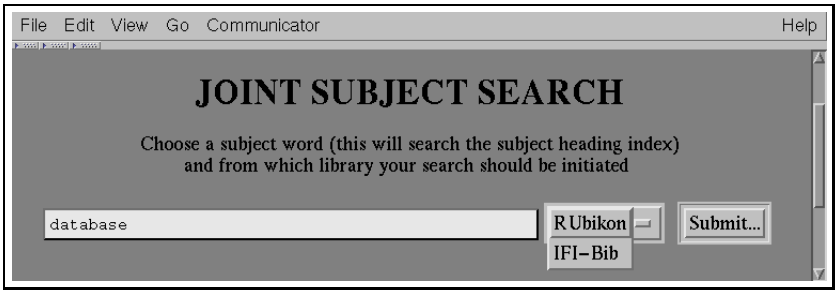

Figure 2: Joint subject search screen

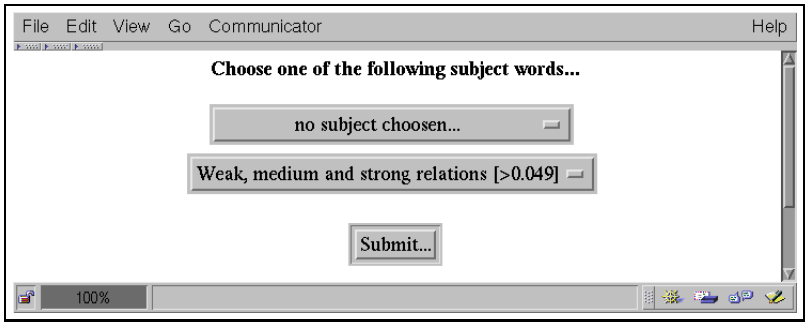

Figure 3: The result screen 1

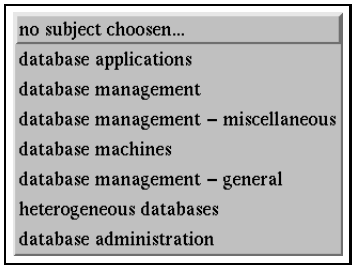

Figure 4: Results screen 2

This will hopefully lead to one or more hits presented in a new pull down menu in the result window (figure 3).

The hits are subject descriptions (with the underlying codes not seen by the user) from the class system chosen in the previous window.

The example here is searching for the word 'database' in the IFI-catalogue. This results in hits in the CRCS system (figure 4). The codes are not shown, only their description.

The user must now choose which aspect of 'database' he wants to follow, and how strong he wants the relations to codes in the second class system to be. We have decided - on the basis of the small material - to provide only three choices here (see figure 5). *Strong, medium* and *weak* refer to specific values of relative frequencies. With a larger material and set of values one could substitute a continuous slider for the pull-down menu in the selection.

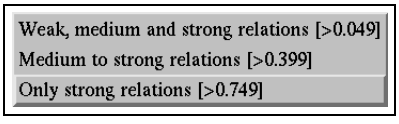

Figure 5: Search options for strength of relations to be displayed.

Figure 6: Search results 3

By submitting the chosen level of relation strength and the chosen aspect we proceed to the next window (figure 6).

In the example we have chosen the broadest set of relevance levels and the term 'database management - general'. The resulting window includes search links to both library catalogues. The first connects to the catalogue of the library chosen in the opening window. The following links connects to the other library and are listed in order of decreasing relation values.

Techniques

The technique behind this general description is based on an inspection of several associative indexes embedded in two Perl programs. Each participating library supplies an index where key and value link **local classification code and a natural language description** of the code[12].

There must also exist a frequency file specifying the **joint classification codes of the different libraries** (or an index with ready-made computations of relative frequencies for all pair of classification codes in accordance with the $Q \rightarrow P$ and $P \rightarrow Q$ formulas, see page 6).

The frequencies of the joint classification codes are based on the joint stock of monographs, as mentioned before. The performance of the programs is pretty good. As long as we work with specific indexes, where the maximum

[12]This index could – as an alternative – contain the total subject index for specific libraries.

number of subject words don't exceede 20000[13] there will be no need for concern - time-wise at least.

3 Discussion

The method used here is a pure quantitative one. It doesn't count for the classification systems used. The systems used could be the same (with different local practice in width and depth) or they could be quite different systems, one universal with a deep hierarchy and the other broad and flat. Our point is that the connection between the codes is based on the actual classification practice.

3.1 The shared titles and the concordance table

The quantitative aspect

This said, the success of this method is of course based on the amount of shared titles, and the number of classification codes used, among other factors.

If two libraries have no or few titles in common the method is useless. This could be the case where two special libraries in different fields meet. If the number of class codes used is also small, the method would produce results with little interest and/or significance.

In our case we found approximately 1000 shared unique titles. We considered this to be less than we wanted, but enough to perform the experiment. In an attempt to boost the amount of joint monographs we considered using 'Universal Standard Bibliographic Code' (USBC) a bibliographic fingerprint method. This was a dead end in our case because the non-ISBN monographs that would be of help were mainly Ph.D theses, student reports and other grey literature. These were not likely to be found in the other library. Thus the effort would not add anything to the data quality.

The 1000 shared titles were classified by 213 different CRCS codes and 347 different UDC codes. This means that as an average every CRCS code is used approximately 5 times, whereas every UDC code is used 3 times. These are very small figures. This means there is a high degree of uncertainty attached to the relation values. It also means we are likely to find one-to-one relations based on only one match of codes. This is why we decided to reduce the importance of low-frequent matches by adding 1 in the denominator of the relation formula.

What we want to say is that we'll have to be careful about drawing too heavy conclusions.

[13] Most thesauri and classification systems like DDC and UDC fulfill this condition.

CRCS code	number of occurences	UDC code	number of occurences	relative frequency
D.1.0	10	681.3.06	9	0.82
		681.3.05	2	0.18
		681.3:159.9	1	0.09
D.1.1	8	681.3.06	8	0.89
		681.3.05	3	0.33
D.1.2	2	681.3.06	2	0.67
		681.3.05	2	0.67
D.1.3	9	681.3.06	7	0.70
		681.3	3	0.30
		681.3.066	2	0.20
		681.3.02	1	0.10
D.1.5	17	681.3.06	12	0.67
		681.3.02	4	0.22
		681.3.05	4	0.22
		681.3.01	3	0.17
D.1.6	4	681.3.06	4	0.80
		681.3.05	3	0.60

Table 4: Enrichment of Universal catalogue by the specialist catalogue

The qualitative aspect

We expected the universal library catalogue to be enriched by the vocabulary from a more detailed description of the computer science library and vice versa. This might seem odd, but - at least in theory - the computer science library would give detailed description of documents in this special field, whereas the universal library would describe non-computer science documents with more detail than the special library.

We have found this to be true, at least in some cases. If we take a look at table 4 we see that every CRCS code in this data set has the UDC code 681.3.06 as its preferred relational code. The description of this code is 'Software', whereas the CRCS codes D.1.* denotes different programming techniques (logic programming, parallel programming, object-oriented programming etc). In the description of the UDC code usage in the table it says among other things "class here programming techniques such as ...". This means that this class code has been enriched with the verbal descriptions of the codes D.1.*.

All in all 681.3.06 occurs 245 times in our material. If we take a closer look at the co-occurences of the UDC 681.3.06 (see table 5) we see that it is even more enriched with terms on software engineering (D.2.*), programming languages (D.3.*) and the theory on meaning and logics of programs (F.3.*).

The end user searching for subjects described by these CRCS codes will be directed by the user interface to the code 681.3.06 when searching the RUb catalogue.

UDC code	number of occurences	CRCS code	number of occurences	relative frequency
681.3.06	245	D.3.2	79	0.32
		D.3.0	19	0.08
		D.3.3	13	0.05
		D.1.5	12	0.05
		D.2.0	10	0.04
		D.2.2	9	0.04
		D.1.0	9	0.04
		D.3.4	9	0.04
		F.3.0	8	0.03
		D.1.1	8	0.03
		F.4.1	7	0.03
		F.3.1	7	0.03
		D.0	7	0.03
		D.1.3	7	0.03
		F.4.0	6	0.02
		D.4.0	5	0.02
		D.2.4	4	0.02
		F.2.2	4	0.02
		D.1.6	4	0.02

Table 5: Co-occurences of the UDC 681.306

If we look behind the pure figures, is this only a matching of codes or is it also a matching of concepts ? If we take a closer look at the description of the class codes we might see how well they relate conceptually (the 2 best)(table 6):

In the reverse table we find that K.4.0 has a relation to 301.152, but it's not very hot. The best matches for K.4.0 are 62:3 (technology and society) and 681.3 (Computers and computer science).

3.2 Updating frequency

A concordance table like this will be dynamic. Changing practices and new shared titles added will have an influence on the relative values over time.

A use of this method will therefore also have to describe an updating routine. Whenever a library updates its classification it must report the ISBN and the codes to a program that checks whether this ISBN is included in the basis of the concordance table. If so, new values have to be computed.

When a library acquires a new ISBN it must report the new ISBN and codes to a transaction table. If the ISBN is already included in the transaction table, the reason is that the other library has acquired this title. Matching ISBNs mean we have got updating material for the concordance table.

If the ISBN is not included, the data is added to the transaction table and residing there until the other library acquires this title.

UDC	CR	Description
301.152.1		MASS MEDIA AND SOCIETY. MASS MEDIA IN GENERAL
	K.4.0	Computers and society - General
	K.4.2	Social issues
681.3:37		COMPUTER SCIENCE. DATA PROCESSING MACHINES : EDUCATION. TEACHING. TRAINING
	K.3.1	Computer uses in education
	K.3.0	Computers and education - General
519.8		OPERATIONAL RESEARCH. THEORY OF GAMES
	G.1.6	Numerical analysis : optimization
	D.4.8	Operating systems : performance
681.3.07		DATA BASE SYSTEMS
	H.2.0	Database management - General
	H.2.1	Database management - Logical design
510.6		MATHEMATICAL LOGIC
	F.4.1	Mathematical logic
	I.2.3	Deduction and theorem proving

Table 6: Intellectual concepts mapped together by statistical analysis

3.3 Another matching possibility

We considered using the Stable Marriage algorithm, but as this algorithm aims at finding one or more one-to-one matching, it wouldn't give us the fuzzyness we needed, the possibility to choose a looser relation (polygamic marriages of both partners).

A refinement of our method would be to iterate over strong relations, removing involved codes from other relations, thereby strengthening the other relation values.

3.4 Further work

Where does this leave us?

We have demonstrated one (of possibly many) methods to overcome the subject classification barrier between libraries. This method is designed to facilitate a natural language approach to subject search which operates on the basis of local classification codes as well as local classification practices unseen from the user point of view.

As a consequence we insist on separating the different libraries in the result / answer screens. This will underline that we are dealing with different libraries and more importantly that we deal with different subject semantics (even though some of the natural language words are "identical").

It is a misunderstanding to believe that a communication protocol like Z39.50 could bridge the gap between library catalogues as regards to sub-

ject searching. The communication protocol operates on another - lower - intellectual level, it is just a means of transporting structured data. Subject search demands coordination on a higher level. This is what our investigation is all about. But of course, we also will need z39.50 as an underlying building block.

In our user interface there will be no need for prior knowledge of the specific classification codes used by the participating libraries (the user will of cause need to know the vocabular of her/his field of interest), and they will be guided to relevant subject codes in the non-local library.

The number of participating libraries in our study are set at an absolut minimum of two, but there is no constraint on this dimension. With a joint server for updating and concordance tables you could easily extend the number of participating libraries to twelve (just to mention a sacred number). Even with an extension like this our user interface will be reasonably easy to use and have a reasonable fast response time.

In a situation where the number of participating libraries exceedss two we will have the benefit of third order relationships where two libraries with little or no record overlap will benefit from concordance with a third library, and thereby establish a possibility for direct search and retrival.

The method we have described could at a general level be used in an attempt to locate subject code clusters in different libraries. We all see (but don't always notice) the 'See also' urge in library catalogues. These requests are a product of an intellectual analysis connecting different but related subject codes, but the classification practice will over a period of time create its own tacit practice. It would be a great help for most subject search practices to have an easy access to these subject code clusters[14] The primary drive in most library search activities is associations. The user is exploring the catalogue. On the one hand searching the known author, the known title. On the other hand the search is aimed at finding the unknown, monographs not even the librarian know existed, material that can provoke to a new understanding, a new experience. We hope that our suggestions will serve this aim.

[14]For a (short) inspirational list of discussions in this area, see:

- Tony Olson & Gary Strawn: *Mapping the LCSH and MeSH systems.* in Information technology and libraries, 1997, p.5–19.
- Mary Micco & Rich Popp: *Improving library subject access (ILSA): a theory of clustering based in classification*, in Library Hi Tech, vol.12, 1994, p.55–66.
- Peter Furniss: *A proposed methodology for examining the provision of subject access in the OPAC*, in International classification, vol.17, 1990, p.85–90.