CEN/TC304 N962 L2/01-022

Subject/Title: Browsing and matching -scoping

Source: PT leader Marc Küster

Date: 9 January (delivered by Project Team 22 December 2000)

Note/Action: Distribution to TC-members for comments. Comments to arrive to Marc Küster (marc.kuester@zdv.uni-tuebingen.de) with a copy to TC/304 secretary (thorgeir@stri.is) before 15th of February 2001, to be reported on in the plenary of TC304 23 February 2001.

Browsing and Matching − scoping

Source: Marc Wilhelm Küster, pt manager

Status: First Draft

Action: Distribution to TC-members for comments. Comments to arrive before February 15th, 2001

Executive summary

Today's society is on its way from a traditionally production-based economy to a knowledge-based economy. The process cannot be stopped.

The European Commission's action plan on *Europe's way to the information society*[1] outlines some of the major developments in this field and recommends steps to be undertaken to prepare Europe for this challenge.

Obviously, the Information Society is not only about *information*, not even only about *access* to information, it is also about *locating* relevant information.

In many ways, information retrieval is the Web revolution's neglected child. Even the otherwise excellent *Information Society Glossary*[2] does not refer to this crucial topic.

Of course, search engines, portal sites, and indexing services do exist. However, in contrast to many of the other topics in this field, the question of locating information involves not only international standards, but also specifically European, national, regional, social, and even personal factors. Many of these issues are related to Europe's multilingual and multicultural heritage which European institutions, including standards bodies such as CEN/TC304 »European localization requirements«, must strive to protect.

The issues encompass points such as:

- Existence of relevant information in many languages;
- The use of different scripts (e. g. Latin, Greek, and Cyrillic scripts);
- The use of letters which are particular to a given language or a number of languages;
- Expectations how such letters or scripts are handled in more restricted character sets such as ASCII (fallback, transliteration, input methods);
- Familiarity with certain cataloguing schemes / database categories specific to a country / a group of countries.

The task soon becomes more ambitious. Human readers[3] will naturally recognize that *sing, sang, sung*[4] are just three tenses of the very same verb, just as *œil* and *yeux* differ only with respect to number. They will also not mix the German word *Boot* with its English homograph of completely different meaning,[5] whereas they understand at once that *Pericles*, *Perikles* and *Περικλῆς* are really one and the same person[6] and that *browsing* and *scanning* can be synonyms[7] in some contexts but not in others.[8]

For English, with its fairly limited number of irregular verbs and its otherwise rather regular construction of derived forms, some of these problems can still be dealt with relatively easily in comparison with most other European languages where word formation is more complex. While no speedy solution is to be expected, these issues must be tackled for the benefit of all non-English speakers in Europe.

---

[1] Cf. `http://europa.eu.int/ISPO/docs/htmlgenerated/i_COM(94)347final.html`
[2] `http://www.ispo.cec.be/infocentre/glossary/i_glossary.html`
[3] Assuming that they are literate in the language(s) in question.
[4] Problem of irregular verb and nouns forms. Declination and conjugation come in here.
[5] Problem of disambiguation.
[6] Problems of non standardized transliteration and of the handling of different scripts. Resolution of spelling ambiguities (e. g. *Göthe* vs. *Goethe*).
[7] Putting to use of thesauri.
[8] Question of matching on natural languages.

Ignoring the European factor is not only contrary to the Commission's stated aim to safeguard Europe's plurality, it also means that European users will be lagging behind in the quest for information.

Scope

Let me begin by quoting the concise terms of reference in CEN/TC304/N739 which specify the project's framework:

> Scope: [...] The objective of this project is to investigate the European needs and problems with searching and browsing, in relation to character sets, transliteration, matching and ordering rules and other cultural specific elements. The needs for a European set of requirements in this area at the present state of technology will be investigated.
>
> Subject and justification: The Global Information Infrastructure must be able to cover European Culturally specific requirements for searching and browsing. Browsing and searching refers to the fast-developing activity around search engines and personal agents operating on large amount of data, implemented mainly as the World Wide Web.
>
> Ultimately, the objective must be that searching and browsing may be carried out in the multilingual environment of Europe.
>
> Technology is moving fast in this area and there are few standards available, although a first generation of products (AltaVista, Lycos, etc.) is available. Consortia such as the W3C or FIPA (Personal Agents) are working in this area. This activity is considered as a key one for GIS (Global Information Society) and one that should see huge developments in the next future.

This study report therefore deals with *European* requirements in the field of *browsing and matching*, the former understood as finding information with the aid of preclassified indexes (organized alphabetically or in a hierarchic structure), the latter as the process of information location in large text corpora, a special case of which is the enormous and ever-changing corpus of the World Wide Web.

It is to be understood that the study focuses on specifically *European* requirements, not on the field of information retrieval *tout court.* Computers have early been used for information storage, and thus, by implication, information retrieval. Unsurprisingly, literature on this topic is sizeable.

From the onset of computing efficient search algorithms have been a core topic of information retrieval and computer science in general.[9] Early on there has also been the desire to transcend the borders of search algorithms and mechanical pattern matching through more intelligent systems that find not only what the user explicitly searches for, but what he *wants* (or rather: may want) to find. Of course, this latter approach is far less concisely defined than the first one, and far more open to cultural − and for that matter personal − expectations. It is here that Europe enters the game.

The study focuses also on browsing and matching of *multilingual corpora*. This is in line with the project's business plan and with the scope of CEN/TC304 which acts as its sponsoring institution.

The study regards corpora which contains data from different historical stages of one and the same language, i. e. diachronic varieties, as a special case of multilingualism.[10] From a technical point of view, the problems are similar, but the willingness of industry to engage itself in this field is often limited as the market may not be big enough to justify large-scale commercial commitment. With the European Commission's aim to offer special support for maintaining Europe's cultural heritage in mind, it is all the more important that this aspect be sufficiently honoured in this report.

[9] The literature on this is almost boundless. (KNUTH, 1973) is often considered *the* classic volume on the subject.

[10] A sample of many of such a project online is the *dokumentasjons prosjektet* (`http://www.dokpro.uio.no/engelsk/index.html`).

The problems that play a rôle here can thus be classified as encompassing diachronic (historical) varieties of a given language as well as coexisting, synchronic varieties of languages. In the latter case, varieties could be dialects within a language, related languages, or altogether unrelated ones.

Overview of the current situation: matching

A brief look at history

As has been stated, research on the problem of searching and pattern matching is old in terms of computer science. In fact, some of the better-known search strategies such as the Soundex method are older than computers.[11]

Pattern matching came into being as a special technique in mechanical translation and automatic language translation,[12] and, while the old optimism that purely mechanical matching techniques are sufficient for translation is long gone, pattern matching has remained a core discipline ever since.

Efficient algorithms are of obvious interest as far as searching and pattern matching are concerned, and have been a constant topic of research. This is, however, well outside the scope of this study report.

Searching was, of course, not always used only on raw text, but early on also for data bases, i. e. data organized into repetitive records of a number of keys each.

Even in the mid seventies data bases could be fairly sizeable − some samples running in the region of 10 GB.[13] However, little attention was given to the design of data retrieval interfaces[14] and user expectations, at least as long as these were not those of an average American.[15] Matching queries were used on the assumption that comparison at a binary level suffices[16] or, at best, that simple case-folding strategies is all that is needed.

Even assuming that no culturally correct matching is intended, the number of different encoding schemes which are in use in Europe[17] makes binary comparison hazardous.

Matching, encodings, and the Universal Character Set (UCS)

The advent of ISO/IEC 10646–1 / Unicode (henceforth: UCS) has to a large degree solved the problem of encoding the languages of Europe in future information pools, though not, of course, of the vast amount of legacy data which could, in principle, be represented in the UCS, though it is unlikely that the conversion will actually take place in the near future.

The UCS has, however, brought problems of its own which are due to the fact that visually and semantically identical characters can be encoded in a variety of ways: For example, the lowercase e with acute (é) might be encoded as `U00E9` or, alternatively, as

---

[11] The patents were registered in 1918 / 1922 (cf. (KNUTH, 1973), p. 391). Paradoxically, many cutting-edge search engines today to not reach that level of sophistication, even though much better search strategies, often dictionary-based, exist nowadays.

[12] Cf. e. g. (LUKJANOW, 1958) und (SALTON, 1966).

[13] The data of the US census was a »large data base [with] approximately 10^{11} bits«, (WELDON, 1975), p. 589. By today's standards, 10 GB would in many fields of application, e. g. full-text data bases, still be fairly large, though some sectors of the industry such as insurance companies and consumer packaged goods companies routinely work with terabyte data bases now.

[14] (GEY, 1975), p. 579, does depict the »casual user« − nicely as woman's hand with coloured finger tips and bracelet.

[15] 15 years later (LI, 1991) still faces the same problem, though all he is asking for is consistency in the user interface.

[16] Cf. e. g. (BURKHARD, 1975), p. 523–525.

[17] Cf. the *Guide on character sets*, `http://www.stri.is/TC304/guide.html`

an e plus the combining diacritic acute, i. e. as the sequence `U0065 + U0301`.[18] Obviously, a user would want to find both forms, if he or she typed the é into a web form.

The W3C Consortium[19] tackles this problem in a technical report on the »Requirements for String Identity Matching and String Indexing«,[20] currently under development. It postulates that »[t]he string identity matching specification shall not expose invisible encoding differences to the user«[21] − a seemingly obvious claim that is not met by most search engines, especially not if we include different encoding schemes.

Not all of the requirements in the report may, however, be in line with European localization requirements. It is highly desirable that European input be given for this report to safeguard European interests in the critical phase of development. A suitable liaison arrangement is to be found.

This already leads us to first demands for action on Browsing and Matching in Europe:

- European cooperation in the development of the technical report on the »Requirements for String Identity Matching and String Indexing«;
- full implementation of the »Requirements for String Identity Matching and String Indexing« must be a top priority once it is in full accordance with European requirements. Furthermore, its guiding principles must be extended to all major encoding schemes in Europe. In terms of working time this is a major task. A guide that fully analyzes these problems would take at least 30 man-days for an encoding expert. It would have to study the topic in terms of a conceptual transformation between legacy character sets and the UCS and identify potential issues such as characters which are missing from the UCS and problems of round-trip convertability.[22]
- In conjunction with this a study must be undertaken on the various encoding schemes in which data is stored in Europe and on the amount of data in the various encoding schemes relative to each other. This study must keep the need for culturally correct matching in focus.[23] This can be estimated as taking at least 30 man-days.

The multilingual approach in today's search engines

The somewhat optimistic assumption that pure pattern matching is enough for culturally correct searching is still more alive than most users would be inclined to assume. While some modern data bases do support multilingual queries, many do not, and even international web search engines such as Lycos and Altavista have but rudimentary

---

[18] More generally, this problem is known as the problem of *canonical equivalence*.

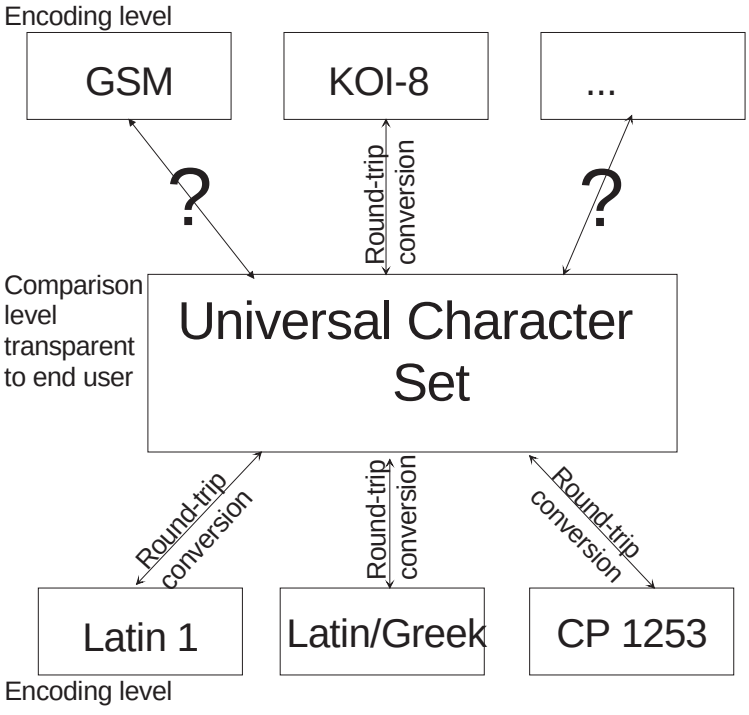
[19] `http://www.w3.org`

[20] `http://www.w3.org/TR/WD-charreq`

[21] Section 2.3 of the TR.

[22] Some common character sets such as the GSM character set which is used in European mobile phones unify characters according to their appearance (they would thus better be termed »glyph standards«). The glyph »A« can thus stand both for the `LATIN CAPITAL LETTER A` and the `GREEK CAPITAL LETTER ALPHA`. This makes it basically impossible to define a unique mapping to and from the UCS.

[23] It might, e. g., be not a top priority to apply intelligent fuzzy search to data that is stored in 5– and 6–bit encoding schemes that support only uppercase letters. On the other hand, certain retrieval requirements such as matching fallback versions of names with the correct spelling, might even be especially relevant in this environment.



Transparent handling of encoding differences for comparison

internationalization support.

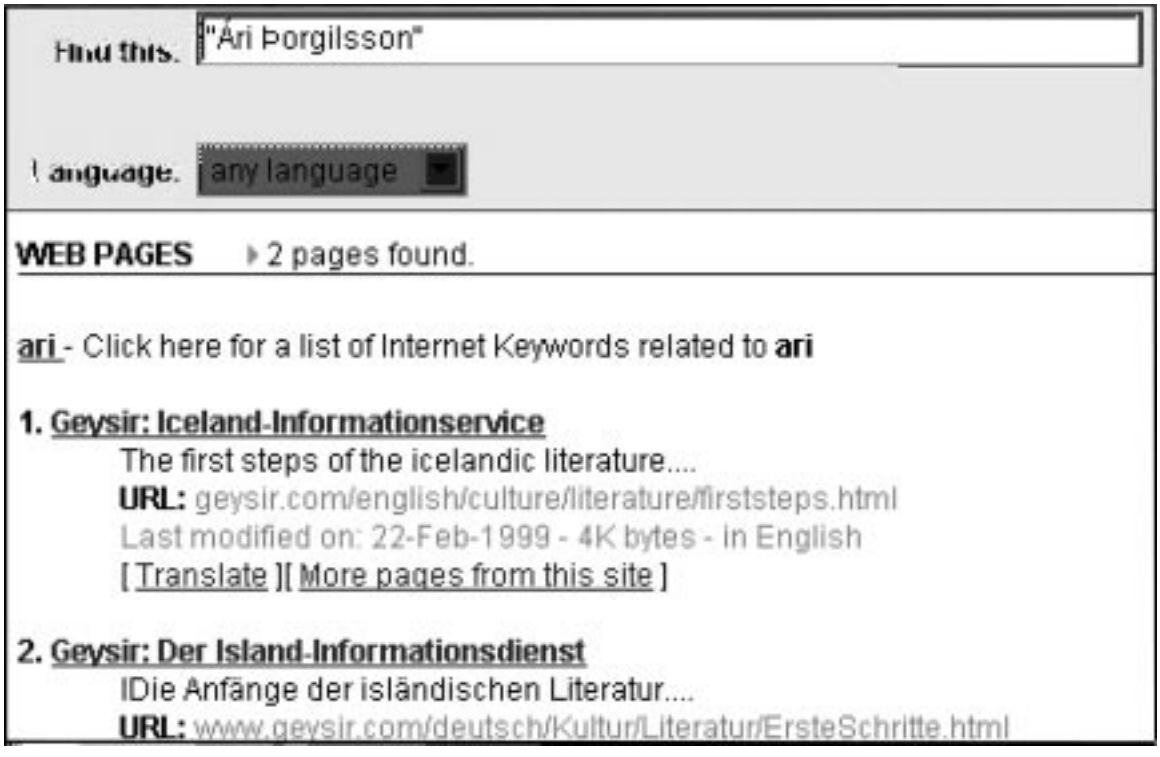
Most search engines do offer a search by language, but few make optimal use of the potential of a consistently multilingual approach.

Let me illustrate this statement with two searches for the famous old Icelandic poet Ári Þorgilsson. The first search with Altavista[24] looks for documents containing his name in the usual fallback spelling Ari Thorgilsson. It finds no less than 44 documents. Many of

these are general descriptions of old Icelandic literature that are highly pertinent to the issue at hand.

[24] The choice of Altavista is by no means menat to denigrate this otherwise excellent service. It is but a sample that would give comparable results with most competitors.

The second try uses the poet's correct name, Ári Þorgilsson. Now the number of hits is only two. There is no overlap between the two hitlists.[25]



Even though this is a very simple and well-known case, the results are markedly different, as the search engine fails to take note of the equivalences Þ/Th and ð/d. In the case of the correct spelling the search fails to deliver most of the relevant hits, rendering this particular search basically useless − and, even worse, actually misleading.

This is all the more true for complex tasks involving, e. g., transliteration between scripts and different established spellings of names.

Most of these are problems which are well-known to library science, though its solutions may not be directly applicable to the IT sector.

Most approaches to deal with these problems are also well-established (and well-entrenched) in the library sector, but differ considerably between European states. One of the more popular schemes are the German *Regeln für die alphabetische Katalogisierung in wissenschaftlichen Bibliotheken* (RAK-WB) which are constantly updated.[26] In the RAK-WB, so called *Ansatzformen* (standard spellings) are prescribed for many of the more important historical names and terms which tend to differ across cultures and time.[27]

It is crucial that the wealth of information and experience which is already available in this and other traditional formats be evaluated for their applicability in Web and database environments, and that suitable implementation guidelines be written.

---

[25] A search for only Þorgilsson / Thorgilsson finds the same documents.
[26] For a list of (amongst others) British cataloguing schemes cf. (ROWLEY, 1992).
[27] Cf. the Pericles, Perikles and Περικλῆς sample, all of which are normalized by (DEUTSCHES BIBLIOTHEKSINSTITUT, 1998), §328, to Pericles, the form used in Latin (!).

Many issues concerning this are collected in the excellent DESIRE Information Gateways Handbook.[28] Nevertheless, this task remains formidable and could be estimated to take in the order of magnitude of 100 man-days.

Completeness of information

Another point of obvious relevance here is the question of *completeness* of the indexes of search engines which a European user may need to access. Data that is not indexed by major search engines will be extremely difficult to locate for the end user, even if the problems above were remedied.

Research on this topic has been undertaken by, amongst others, the Working Group of the IRT (Internet Retrieval Tools). A preliminary report in Dutch is available.[29] On the basis of 11 popular search engines it monitors systematically if and, if so, with which time lag information − in the concrete case a small Dutch text − is indexed. It also points out many problems in a truly multilingual environment, as indexing works often less than ideal for texts which are not in ISO/IEC 8859–1. Even for data in that popular character set, problems with different storage formats for letters with diacritics − e. g. *Méditerranée* can also be stored as *Méditerranée* − causes problems for certain search engines.

The research of the IRT should be supported and the results given wider publicity. Special focus should be given to the behaviour of search engines with respect to letters with diacritics. Recurrent reports on this topic could heighten customer awareness of these annoying shortcomings and show that solutions are possible. Customer demands would give an incentive to industry to support European requirements better.

Linguistically aware matching

Linguistically aware matching in its widest sense encompasses all matching strategies that exploit information on the phonetic, syntactic, and semantic properties of a given language. In this understanding it coincides with important fields of study in computer linguistics and is too generic for scoping in this report.

This study shall restrict the definition, for the time being, to strategies that function on the word formation level, thus contrasting it with *thesauri* which try to evaluate synonyms and near-synonyms on a semantic level.[30]

For all inflecting languages − the great majority of languages spoken in Europe − the problem here is that of locating not only the search term itself, but also its inflected forms. For English, the solution is still fairly straightforward and can be handled with some degree of success via substring matching (matching on truncated strings). In this manner, a search for *match* finds also inflected forms such as *matching* or *matched* (not vice-versa, of course).[31]

[28] `http://www.desire.org/handbook`

[29] Cf. (VAN DER LAAN, 1999).

[30] A thesaurus is usually defined as »a controlled vocabulary of semantically and genetically related terms covering a specific area of knowledge« ((PAO, 1989), p. 119).

[31] There are, of course, many problem cases even in English where such simple way forward does not succeed, e. g. the irregular verbs.

For many other European languages, this procedure does not work at all. Thus, a substring search for the German word stem *find* does locate the infinitive *finden*, but neither the past *fand*, nor the past participle *gefunden* nor many of the composites such as *Fundstelle*. In this case, it is often considered necessary to use dictionaries to reduce both the search expressions and the target data to a standard form.

Historically, many of these problems originate in the old Indo-European *Ablautreihen*. The original Indo-European mother language could form semantically related forms by changing or omitting vowels of the word stem. Since most European languages are of Indo-European origin − notable exceptions are the Finno-Ugric languages, the Turkic languages (related to the former), various Caucasian languages and Basque −, this is a common feature that it might be possible to exploit to develop generic tools for this language family. It can be an interesting project for a team of specialists in historical linguistics and in information retrieval to explore the possibilities of an approach that focuses on parallel structures in all Indo-European languages. A first analysis of the situation with proof of concept implementations or the result that this approach is not fruitful could be achieved in around 40–50 man-days. A similar project with a slightly more agressive time frame (25 man-days) could be envisaged for the Finno-Ugric languages.

In contrast to this another project should evaluate the efficiency of dictionary based approaches. It should perform a basic market study of products for the major European languages that use this approach and contrast them. A project could perform such a task in around 20 man-days.

In the long term, it should be studied if and how dictionary-based and structure-based approaches could be entwined for maximum efficiency.

There are more possible projects that could be fruitful in this field of action, both on European requirements and on ongoing research. The project team recognizes that within the schedule and the time constraints, even this process of scoping can only be preliminary and a first step towards a larger project.

Phonetically aware matching

Though logically a subset of linguistically aware matching, phonetically aware matching is here treated separately. Although still complex enough, it is in comparison a more straightforward task where a number of products has already hit the market − at least for the English language.

Some of the earliest techniques in this field, such as the fairly simplistic Soundex method which tries to mirror any given spelling of an English word to what it considers its phonetic skeleton, were developed well before the advent of the computer, let alone the internet.

Nowadays, many commercial products such as the *Encyclopædia Britannica* database engine feature phonetically aware matching which, apart from the phonetic structure, also tries to accommodate common spelling errors. For the English language, the results seem to be fairly satisfactory.

For languages other than English some of the methods such as Soundex fail to give satisfactory results, as the rules are ill-adapted to the phonetic structure of the languages in

question. The relationship between spelling and pronunciation is highly language-dependent. Field experiments with the TUSTEP-based *Online Public Access Catalogue (OPAC)* of the University of Tübingen's computing centre[32] have revealed that even for its relatively small database of some 60.000 items Soundex delivers unacceptably many false hits.

It is desirable that a study be undertaken that lists and evaluates all European projects and products (both commercial and academic) in this field and compiles a status report. This study should then proceed to point out which European requirements are not yet met and give guidance on how shortcomings can be remedied.

In contrast to the whole of *linguistically aware matching,* this study could be accomplished in a reasonable timescale if it is restricted to the state languages of the CEN countries (phase one). It should be realistic to complete the study in 30–40 man-days.

Thesauri and the problem of disambiguation

Ideally, a search engine should give assistance also on a semantic level. Terms like *searching* and *matching* or *hit* and *match* might be thought of as synonyms in certain circumstances, though not in others. A user who searches for one of these terms might want to locate documents on the others also.

Furthermore, a user might also want to find documents in other languages than the one the query was formulated in. In this case, he would want not only synonyms, but also translations of the original search expression.

For such a mechanism to work, the search term needs to be disambiguated first. Otherwise, the end user will be confronted with results which are based on wrong equivalences.

For the time being at least, both functionalities would have to be user-configurable to allow the user to avoid looking for synonyms at all or to exclude certain unsuitable synonyms from the thesaurus. For translations this is even more important, as a user might not be interested in documents in languages which he or she cannot read, though automatic translation services such as envisaged by the international UNL-project[33] might in the foreseeable future alleviate that problem.

---

[32] `http://www.uni-tuebingen.de/cgi-bin/zdvlit`

[33] For more information cf. `http://www.iai.uni-sb.de/UNL/unl-en.html`. UNL stands for Universal Network Language, a language-independent metasyntax that allows for easy translation between major world languages.

Overview of the current situation: browsing

Background information

If matching allows automated access to information via a query which the user submits, browsing assumes a pre-defined structure in which the user selects a concept either alphabetically[34] or by descending through a hierarchic structure, the latter being the usually preferred way for large databases.[35]

Indexing is in this context the creation of short condensed descriptions of the documents or sites and directories that can identify the content of the source information without actually having to consult the source itself. It is often combined with a pure browsing approach to allow the user to preselect items of interest.

Browsing as a concept is again much older than computing. Most freely accessible libraries function along these lines: books are arranged first by very general terms (say, mathematics, philology, philosophy, ...) and then by subsequently more specialized ones (say, analysis, Latin, Platonism, ...). A user can then walk by the shelves and look for the titles which pertain to his or her field of interest.

In some countries such as Switzerland catalogues must be multilingual even on a national level; again, libraries had to face this problem quite early and undertook efforts to get to terms with this. For example, between 1997 and 1999 the *CoBRA+ working group on multilingual subject access*[36] tried to establish a trilingual list of corresponding subject headings (in German, French and English). Work goes on in projects such as *MUSE*, the *Multilingual Subject Entry*, a cooperation between the national libraries of Switzerland, Germany, France, and Great Britain. Many issues remain, e. g. the cases where there are no clear equivalencies between the lists. A concept in the source list may correspond to several ones in the target list or on the contrary have no obvious equivalent at all. Some quite usual headings in one language may also coincide with a whole cluster of headings in another one.

On the Web, the first browsing applications started as simple link lists where a user had amassed all information he or she could find on a favourite subject. Over time, some of these became larger, more varied in subject matter and were renamed into portal sites.[37]

Nowadays, both browsing and matching approaches are found in both large-scale commercial applications such as Yahoo![38] or, for Germany, DINO,[39] and in academic endeavours such as the renowned Gnomon project.[40]

---

[34] Alphabetic lists are often used to list indexes of various kinds in aid of search engines. A classical case is an OPAC which allows for search of the author name, but offers also an author index with the chosen cataloguing forms or a list of keywords. For an exemplary discussion of some of the problems cf. also (MURPHY, 1991), section 7.10.

[35] For an elegant graphic juxtaposition between browsing and matching (here called querying) cf. e. g. (MIKOLAJUK, 1991), p. 86f.

[36] Cf. also `http://portico.bl.uk/gabriel/cobra/finrap3.html`

[37] This kind of service is also known as a *subject based information gateway* (SBIG).

[38] `http://www.yahoo.com`

[39] `http://www.dino-online.de`

[40] `http://www.gnomon.ku-eichstaett.de/Gnomon/`

Internationalization and localization issues

A good choice of categories is the key to success for a portal site. However, a sensible choice of categories is by no means culture-independent. Many issues are of great importance in one country or culture, but quite irrelevant to another − especially, when they touch upon political, cultural and educational issues and corresponding institutions. A good example are institutions such as colleges that are omnipresent in countries like the US and Japan, but basically non-existent in many European countries.

Some of the indexing services such as Yahoo try to take this into account. While they attempt to present as unified an interface as possible across countries and maintain a consistent set of primary categories, the subcategories differ significantly. This can be evinced by looking at the keywords that Yahoo selects for the primary category *Education* and their national equivalents for France, Germany and the US as of October 2000:[41]

France: Actualités et médias (83), Annuaires et guides Web (24), Chaînes de sites (2), Conseil et orientation (18), Emploi (29), Enseignement primaire (51), Enseignement professionnel (243), Enseignement religieux@, Enseignement secondaire (113), Enseignement spécialisé (23), Enseignement supérieur (767), Espaces de discussion (1), Examens, concours et diplômes (16), Logiciels (6), Organismes (196), Pédagogie (175), Ressources pour l'enseignement@, Salons, colloques et conférences (15), Sociétés@, Technologies d'enseignement (62).
Germany: Behinderte (39), Berufseinstieg (182), Bibliotheken@, Bildungspolitik (60), Bücher@, Erwachsenenbildung und Weiterbildung (401), Fernunterricht (40), Finanzielle Unterstützung (15), Firmen@, Forschung (242), Frauen@, Hilfestellung (20), Hochbegabung (20), Hochschulen (1165), Institute (122), Konferenzen und Messen (7), Lesben, Schwule und Bisexuelle@, Nach Fachgebiet (16), Nachrichten und Medien (68), Organisationen (19), Pädagogik (208), Programme (73), Schulen (294), Schulwesen (801), Verzeichnisse (33).
US: Academic Competitions (76), Adult and Continuing Education (316), Bibliographies (5), Bilingual (24), Career and Vocational (232), Chats and Forums (43), Companies@, Conferences (46), Correctional@, Disabilities@, Distance Learning (475), Early Childhood Education (81), Employment (133), Equity (26), Financial Aid (377), Government Agencies (76), Graduation (58), Higher Education (18217), Instructional Technology (322), Journals (31), K–12 (52762), Literacy (10), News and Media (83), Organizations (2903), Policy (50), Programs (288), Reform (63), Special Education (163), Standards and Testing (61), Statistics (6), Teaching (90), Theory and Methods (625), Web Directories (43).

Even without extensive knowledge of the different European educational systems and their contrast to the US one notices that they are directly reflected in the choice of keywords even beyond the »hard facts«: Frequently the understanding of a given concept such as education differs significantly from culture to culture.[42]

The often widely divergent understanding of seemingly obvious concepts poses some of the most difficult problems for internationalization and localization as these topics are in their nature imprecisely defined and thus ill-suited for formalization into clearly applicable guidelines, let alone hard and fast rules. Many such issues are even within one culture the subject of dissension and may thus have the potential to unwittingly antagonize a certain percentage of the customer base and / or be the origin of unfavourable comments in the media.

---

[41] The numbers in brackets give the number of relevant hits. The @ indicates a link into a different system of categories.
[42] In principle, this is a special case of the issues that arise when translating thesauri.

While the list of keywords in education are in themselves an interesting comment on cultural conventions, significant cultural differences between the CEN countries make it difficult to see how a pan-European approach could be designed that does not violate the principle of subsidarity. Yet, a compromise is needed as soon as a pan-European organization aims to offer a broad, multilingual portal site. CEN/TC304 must look into this matter and liaise with such organizations to perform a market study of what is actually planned in this respect. If requirements for pan-European portals can be demonstrated to exist, CEN/TC304 must set up a project team that works in cooperation with commercial vendors on such a categorization.

Manual intervention

Unlike the automatic brute-force indexing, the categorization of links normally requires extensive human intervention.[43] Admittedly, for texts which belong to a well-defined category with a fairly well-defined vocabulary, e. g. research papers in engineering, it is possible to get fairly far with autmomatic classification.[44] It is of considerable interest to explore this path towards automatization in greater detail and to find out for which kind of documents it could equally well be applied.

While the main thrust towards new companies is likely to come from the private sector, it is worthwhile to seek out the well-defined areas of public interest where automatic classification techniques can be put to best use.

In the United States startup companies exist which put such approaches to use for, e. g., the automatic collection and classification of job openings.[45] For Europe, this is of specific interest, as a unified labour market can be furthered by making it easier to find jobs online also outside of an applicants country of origin. While there are a considerable number of both public and commercial agencies of this kind already on the web for national labour market, their pan-European linkage is especially for the public sector not always a reality.

Automation is helpful in certain, highly standardized areas. In most cases, however, the contents of a document must be read − ideally by a person with a certain expertise in the topic concerned − and then suitable metadata must be added which determines its place in the hierarchic structure. Unlike books, which can only reside in one place, electronic documents can be assigned to several positions, if their contents warrant this.

Similar strategies were considered in the early 90s for OPACs − once more libraries played the rôle of a forerunner. Suggestions such as keying in the table of contents as a book's abstract and to make use of this information to create »subject clusters«[46] that should allow users to browse by topic, were explored at places such as the Library of Congress. Similar concepts were implemented for the Web-environment via the `META` tag mechanism which was intended as a means of the page's author to provide keywords. Unfortunately, this mechanism was subjected to widespread misuse by people who tried to draw people to their pages by inserting misleading information. For this reason, this mechanism is increasingly falling out of use again.[47]

[43] An enterprise such as Yahoo! occupies a large part of its workforce just for reading and cataloguing web sites.

[44] For an example cf. `http://cora.whizbang.com`. This collection of research papers has been constructed using machine learning strategies.

[45] An example of a company that uses such a strategy is `http://www.flipdog.com`

[46] (MICCO, 1991), p. 129.

Approaches such as the Dublin Core[48], a »metadata format defined on the basis of international consensus which has defined a minimal information resource description, generally for use in a WWW environment«, can suffer from similar drawbacks if not applied with the necessary stringency − a fact, that the initiators of the Dublin Core are well aware of. The main problem is the consistent use of cataloguing strategies in an often decentralized and not always professional environment. Furthermore, even such popular metadata schemes as the Dublin Core are not fully internationalized and fail to fulfil pan-European requirements. It is desirable to participate in the ongoing internationalization endeavours of the Dublin Core »Multiple Languages« working group[49] to ensure that European requirements are met. This project must involve European research libraries. It could be estimated at around 20 man-days.

In sum: Human intervention is at the same time the asset and the drawback of the browsing approach. On the one hand, a well-made portal site can offer a level of service to the end user that a brute-force search engine cannot (and, in the foreseeable future, will not be able to) deliver. On the other hand, the need for manual intervention means that it cannot be as extensive in coverage and as speedy in reaction as a web crawler.

Indexing services

State of the art indexing services such as are planned by the *Pilot Index Service for Research and Education in Europe*, short *REIS – Pilot*,[50] now complemented by the *TERENA Portal Coordination Workshop*, intend to classify Europe's wealth of multilingual Web information (estimated at some 20–30 million pages) using manual and automated classification tools. The resulting indexes should be both searchable and browsable by subject, thus functioning not only as a value-added search engine, but also as a portal site. Such an index repository will reflect Europe's multi-subject, multilingual, cross-border, and multi-cultural data online.

The complexity of the information, the fact that no single place can assemble the required expertise in languages and subject matters, makes it evident that any such approach must by necessity be working in a distributed mode.

In a first phase, specific communities must be targeted rather than the complete web. This can include the building of community oriented portal services, cross-language searching, searching for authoritative sources of information, creating new communities of trust, establishing reliable methods for assigning metadata etc.

Projects such as REIS may also serve as contact places for the technical coordination of many of the projects which are suggested in this study report.

The Holy Grail

The ideal world would, of course, combine the best of both approaches and offer a browsable subject index that would be automatically culled from the web itself. Extensive

[47] Cf. also (VAN DER LAAN, 1999), section »Header«, on an overview of current practice in this field.
[48] http://www.purl.org/DC
[49] http://www.purl.org/DC/groups/languages.htm
[50] http://www.terena.nl/projects/reis/

research is going on in that direction, e. g. at the Swedish Institute for Computer Science (SICS)[51] in Sweden. While some preliminary results are published, a long way still remains to be gone before this research may one day mature into products that are viable on the market.

The amount of work still needed for this kind of research cannot be estimated within this scoping report. It is, however, evident that Europe has a massive interest in the successful conclusion of such developments.

European requirements

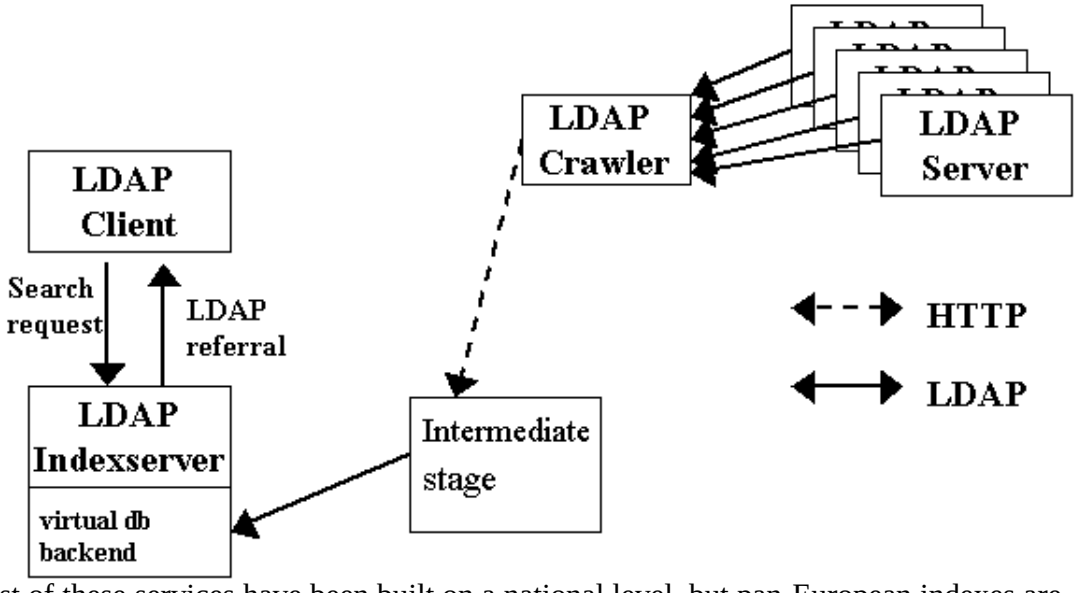
For some languages and subjects, reasonably well-working lists have been compiled and are maintained by either commercial enterprises or academic institutions. It would be highly desirable to compile a »list of lists« which catalogues existing portal sites by European language, ideally giving also some idea of the coverage of the portal site in question. Here, some groundwork was done by the portal sites themselves, but a lot still needs to be done. This effort would at the same time point out which languages are, as yet, poorly served in this regard and would give an incentive to build such services there also.

It is realistic that a survey of the market could be undertaken in around 20–30 man-days. The deliverable would, in this case, be web-based as a matter of course. The main problem would be to find a maintenance agency that ensures that the catalogue stays up-to-date.

[51] http://www.sics.se

Digression: Common Indexing Protocol implementations

Implementations of the Common Indexing Protocol (CIP) aim at establishing indexing services for directories of certain, usually highly formalized entities. Examples might include phone books (White Pages), Yellow Pages and e-mail directories. CIP implementations work in a highly distributed mode. Their indexes are based on local data collections, often one collection per institution. At the same time it allows the end user to transparently access all directories on a national or multi-national level through a *referral service*, basically a crawler-built index that refers the search client to the correct local data collection.[52]



Most of these services have been built on a national level, but pan-European indexes are urgently needed to facilitate personal and commercial contacts across national borders.[53] While these indexes have usually only few categories such as the person's name, company affiliation, address, and phone number per entry, the contents of these categories differ quite drastically even through Europe. E. g., a personal name consists in most parts of Western Europe of one or more first names followed by the family name. In Russia, however, it is formed by a first name followed by father's name and finally family name, whereas in most cases in Iceland it comprises only first name and father's name. In addition to these difficulties, a person may have one or more titles or academic degrees, the rules for the handling of which is once more not unified across Europe.

Differences such as these are mirrored in the potentially incompatible contents for seemingly identical categories which in turn endanger interoperability between the various services which exist on a national level. Problems such as these are currently faced by state of the art projects such as the *task force for LDAP Service Deployment (TF-LSD)*[54], and with implementations such as the *Global Indexing Directory Service* (GIDS)[55].

[52] For the illustration I would like to thank my colleague Peter Gietz.
[53] A *proof of concept* implementation has been made by the DESIRE II project. Cf. `http://www.desire.org`
[54] `http://www.terena.nl/task-forces/tf-lsd`

To enable pan-European and gloabal indexing directory systems, the concrete internationalization issues in current CIP implementations must be scrutinized and suitable recommendations for clean, localizable architectures must be formulated in cooperation with leading projects. This task can be estimated at around 15 to 20 man-days.

[55] Cf. `http://gids.catalogix.se/gids.html`

Table: List of proposed projects

To be added as soon as the list stabilizes. A tentative priority of projects may also be added here.

References

Internal to standardization

- CEN/TC304 N739: Terms of reference for P27,1 European matching rules: (scoping);
- CEN/TC304 N752: Terms of reference for P27,2 European matching rules;
- CEN/TC304 N780: General rules for Project Teams;
- CEN/TC304 N785: Call for experts;
- CEN/TC304/N860: Business plan of pt Matching;
- CEN/TC304/N924: Project manager's progress report on Browsing and Matching (scoping);
- CEN/TC304/N925: Browsing and Matching (scoping): Pre-Draft 1.1;
- CEN/TC304/NXXX: Browsing and Matching (scoping): Disposition of Comments on Pre-Draft 1.1;
- World Wide Web Consortium Working Draft 10–July–1998: Requirements for String Identity Matching and String Indexing[56]
- Draft Unicode TR 15: Unicode Normalization Forms[57];
- Multi-lingual issue − DESIRE Information Gateways Handbook[58].

External to standardization

(BURKHARD, 1975) Burkhard, Walter A.: Partial-match queries and file designs. In: (KERR, 1975) p. 523 –525.

(DILLON, 1992) Dillon, Martin (ed.): *Interfaces for Information Retrieval and Online Systems. The state of the art.* New York. Greenwood Press: 1992.

(GEY, 1975) Gey, Frederic; Mantei, Marilyn: Keyword access to a mass storage device at the record level. In: (KERR, 1975) p. 572 –588.

(HAYS, 1966) Hays, David G. (ed.): *Readings in automatic language processing.* New York. Elsevier: 1966.

(KERR, 1975) Kerr, Douglas S. (ed.): *Proceedings of the international conference on very large data bases.* New York. ACM: 1975.

(KNUTH, 1973) Knuth, Donald E.: *The art of computer programming. Sorting and searching.* Reading, MA. Addison-Wesley: 1973. Vol.: 3.

(LI, 1991) LI, Tian-Zhu: Generic Approach to CD-ROM Systems: A Formal Analysis of Search Capabilities and Ease of Use. In: (DILLON, 1992) p. 259 –275.

(LUKJANOW, 1958) Lukjanow, Ariadne W.: The C. M. T. (Code Matching Technique) mechanical translation process. In: (WEISS, 1958) p. 60.1 –60.2.

(MICCO, 1991) Micco, Mary: The Next Generation of Online Public Access Catalogs: A New Look at Subject Access Using Hypermedia. In: (TYCKOSON, 1991) p. 103 –132.

(MIKOLAJUK, 1991) Mikolajuk, Zbigniew; Chafetz, Robert: A Domain Knowledge-based, Natural-Language Interface for Bibliographic Information Retrieval. In: (DILLON, 1992) p. 83 –105.

[56] `http://www.w3.org/TR/WD-charreq`
[57] `http://www.unicode.org/unicode/reports/tr15`
[58] `http://www.desire.org/handbook/`

(MURPHY, 1991) Murphy, F. J.; Pollitt, A. S.; White, P. R.: *Matching OPAC User Interfaces to User Needs.* Huddersfield. British Library Research & Development Departement: 1991. Vol.: 6041. In: British Library R & D Report.

(PAO, 1989) Pao, Miranda Lee: *Concepts of information retrieval.* Englewood, Colorado. Libraries Unlimited: 1989.

(DEUTSCHES BIBLIOTHEKSINSTITUT, 1998) Deutsches Bibliotheksinstitut (ed.): *Regeln für die alphabetische Katalogisierugn in wissenschaftlichen Bibliotheken (RAK-WB).* Berlin. DBI: ^{2} 1998.

(ROWLEY, 1992) Rowley, Jennifer E.: *Organizing Knowledge. An Introduction to Information Retrieval.* Aldershot. Ashgate: ^{2} 1992.

(SALTON, 1966) Salton, Gerard A.: Automatic phrase matching. In: (HAYS, 1966) p. 169 –188.

(TYCKOSON, 1991) Tyckoson, David E. (ed.): *Enhancing Access to Information: Designing Catalogs for the 21st Century.* Binghamton. Haworth Press: 1991.

(VAN DER LAAN, 1999) van der Laan, Hans: *De Werkgroep IRT.* To be published. 1999.

(WEISS, 1958) Weiss, Erik A. (ed.): *Preprints of Summaries of Papers Presented at the 13th National Meeting Association for Computing Machinery. Urbana, Illinois June 11–13, 1958.* New York. ACM: 1958.

(WELDON, 1975) Weldon, Jay-Louise: Implementation strategies for the census data base. In: (KERR, 1975) p. 589 –590.