Browsing and Matching − scoping

Source: Marc Wilhelm Küster, pt manager

Status: Pre-Draft 1.1

Action: Distribution to TC-members for comments. Comments to arrive before December 17th, 1999

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.

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.

---

[1] Cf. also `http://www.ispo.cec.be/infosoc/backg/action.html`
[2] `http://www.ispo.cec.be/g7/backg/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

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 latter understood as the process of information location in large text corpora, a special case of which would be 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 explicitely searches for, but what he *wants* (or rather: may want) to find. Of course, this latter approach is far less concisely defined as 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 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 general market relevance is often considered to be not sufficiently great 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.

The problems that play a rôle here can thus be classified into two dimensions:

---

[9] The literature on this is almost boundless. [KNUTH73] 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`)

− temporal: access to texts written in other stages of the user's cultural heritage;
− spatial: access to texts written in other cultures than the user's own.

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, many of the advanced search strategies such as the Soundex method are much 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 large − 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]

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 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 an e plus

[11] The patents were registered in 1918 / 1922 (cf. [KNUTH73] , p. 391). Paradoxically, many cutting-edge search engines today to not reach that level of sophistication
[12] Cf. e. g. [LUKJANOW58] und [SALTON66]
[13] The data of the US census was a »large data base [with] approximately 10^{11} bits«, [WELDON75] , p. 589
[14] [GEY75] , p. 579, does depict the »casual user« − nicely as woman's hand with coloured finger tips and bracelet
[15] 15 years later [LI91] still faces the same problem, though all he is asking for is consistency in the user interface
[16] Cf. e. g. [BURKHARD75] , p. 523–525
[17] Cf. the *Guide on character sets*

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;
- a 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 would be a major task. A guide that fully analyzes these problems would take at least 50 man-days for an encoding expert.
- In conjunction with this a study must be undertaken on the relative availability of data in various encoding schemes and the need for culturally correct matching.[22] This can be taxed at at least 30 man-days.

Trends of today: 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 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 CEN/TC304's secretary, Mr. Þorgeir Sigurðson from STRÌ, Iceland. The first search with Altavista looks for documents containing his name in the usual fallback spelling Thorgeir Sigurdson. It finds but one document.

The second try uses his correct name, Þorgeir Sigurðson, difficult to input from many non-Icelandic keyboards. Now the number of hits is 28, but the list does not include the previously found document.

---

[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] 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.

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 usual equivalences Þ/Th and ð/d.

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.[23] 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.[24]

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.

Many issues concerning this are collected in the excellent DESIRE Information Gateways Handbook.[25] Nevertheless, this task remains formidable and could be taxed 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 indices 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.[26] 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 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 should give an incentive to industry to support European requirements.

Linguistically aware matching

Linguistically aware matching in its widest sense encompasses all matching strategies

---

[23] For a list of (amongst others) British cataloguing schemes cf. [ROWLEY92]

[24] Cf. the Pericles, Perikles and Περικλῆς sample, all of which are normalized by [RAK98] , §328, to Pericles, the form used in Latin (!)

[25] `http://www.ilrt.bris.ac.uk/~ecmb/desire/hb/`

[26] Cf. [VANDERLAAN99]

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.[27]

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).[28]

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 not neither the past participle *gefunden* nor many composites. In this case, it is necessary to use dictionaries to reduce both the search expressions and the target data to a standard form.

The project team still intends to do some scoping on this field of action, both on European requirements and on ongoing research, but it 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 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[29] 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.

---

[27] A thesaurus is usually defined as »a controlled vocabulary of semantically and genetically related terms covering a specific area of knowledge« ([PAO89] , p. 119)

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

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

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. 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[30] might in the foreseeable future alleviate that problem.

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[31] or by descending through a hierarchic structure, the latter being the usually preferred way for large databases.[32]

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.

---

[30] 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

[31] Alphabetic lists are often used to list indices of various kinds in aid of search engines. A classical case would be 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 [MURPHY91] , section 7.10

[32] For an elegant graphic juxtaposition between browsing and matching (here called querying) cf. e. g. [MIKOLAJUK91] , p. 86f

On the Web, the first applications in this direction 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.

Nowadays, both browsing and matching approaches are found in both large-scale commercial applications such as Yahoo![33] or, for Germany, DINO,[34] and in academic endeavours such as the renowned Gnomon project.[35]

Unlike the automatic brute-force indexing, the categorization of links normally requires extensive human intervention.[36] The contents of a document must be read − ideally by a person with a certain expertise in the topic concerned − and then be assigned to its place in the hierarchic structure. Unlike books, which must reside in one place, electronic documents can be assigned to several positions, if their content warrants 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«[37] 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.[38]

Approaches such as the Dublin Core, a set of XML attributes, can suffer from similar drawbacks if not applied with the necessary stringency. 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 would be desirable to participate the ongoing internationalization endeavours of the Dublin Core »Multiple Languages« working group[39] to ensure that European requirements are met. Such a project could be taxed at around 20 man-days.

Summary

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.

---

[33] `http://www.yahoo.com`
[34] `http://www.dino-online.de`
[35] `http://www.gnomon.ku-eichstaett.de/Gnomon/`
[36] An enterprise such as Yahoo! occupies a large part of its workforce just for reading and cataloguing web sites
[37] [MICCO91] , p. 129
[38] Cf. also [VANDERLAAN99] , section »Header«, on an overview of current practice in this field
[39] `http://www.purl.org/DC/groups/languages.htm`

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*,[40] 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 indices 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.

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 research is going on in that direction, e. g. at the Swedish Institute for Computer Science (SICS)[41] 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.

This kind of research is not taxable within a 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 lists the major indices by European language. 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.

Table: List of proposed projects

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

---

[40] `http://www.terena.nl/projects/reis`
[41] `http://www.sics.se`

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/NXXX: Project manager's progress report on Browsing and Matching (scoping);
- World Wide Web Consortium Working Draft 10–July–1998: Requirements for String Identity Matching and String Indexing[42]
- Draft Unicode TR 15: Unicode Normalization Forms[43];
- Multi-lingual issue − DESIRE Information Gateways Handbook[44].

External to standardization

[BURKHARD75] Burkhard, Walter A.: *Partial-match queries and file designs.* In: [KERR75] p. 523–525. 1975.

[DILLON91] : Dillon, Martin (ed.): *Interfaces for Information Retrieval and Online Systems. The state of the art* Greenwood Press: New York. 1992.

[GEY75] Gey, Frederic; Mantei, Marilyn: *Keyword access to a mass storage device at the record level.* In: [KERR75] p. 572–588. 1975.

[HAYS66] Hays, David G. (ed.): *Readings in automatic language processing.* Elsevier: New York. 1966.

[KERR75] Kerr, Douglas S. (ed.): *Proceedings of the international conference on very large data bases.* ACM: New York. 1975.

[KNUTH73] Knuth, Donald E.: *The art of computer programming. Sorting and searching* Addison-Wesley: Reading, MA. Vol.: 3. 1973.

[LI91] LI, Tian-Zhu: *Generic Approach to CD-ROM Systems: A Formal Analysis of Search Capabilities and Ease of Use.* . In: [DILLON91] p. 259–275. 1991.

[LUKJANOW58] Lukjanow, Ariadne W.: *The C. M. T. (Code Matching Technique) mechanical translation process.* In: [WEISS58] p. 60.1–60.2. 1958.

[MICCO91] Micco, Mary: *The Next Generation of Online Public Access Catalogs: A New Look at Subject Access Using Hypermedia.* . In: [TYCKOSON91] p. 103–132. 1991.

[MIKOLAJUK91] Mikolajuk, Zbigniew; Chafetz, Robert: *A Domain Knowledge-based, Natural-Language Interface for Bibliographic Information Retrieval.* . In: [DILLON91] p. 83–105. 1991.

[MURPHY91] Murphy, F. J.; Pollitt, A. S.; White, P. R.: *Matching OPAC User Interfaces to User Needs.* British Library Research & Development Departement: Huddersfield. Vol.: 6041. In: *British Library R & D Report.' 1991.*

[42] `http://www.w3.org/TR/WD-charreq`
[43] `http://www.unicode.org/unicode/reports/tr15`
[44] `http://www.ilrt.bris.ac.uk/~ecmb/desire/hb/2-13.html`

[PAO89] Pao, Miranda Lee: *Concepts of information retrieval.* Libraries Unlimited: Englewood, Colorado. 1989.

[RAK98] : Deutsches Bibliotheksinstitut (ed.): *Regeln für die alphabetische Katalogisierugn in wissenschaftlichen Bibliotheken (RAK-WB).* DBI: Berlin. 1998.

[ROWLEY92] Rowley, Jennifer E.: *Organizing Knowledge. An Introduction to Information Retrieval* Ashgate: Aldershot. 1992.

[SALTON66] Salton, Gerard A.: *Automatic phrase matching.* In: [HAYS66] p. 169–188. 1966.

[TYCKOSON91] : Tyckoson, David E. (ed.): *Enhancing Access to Information: Designing Catalogs for the 21st Century*. Haworth Press: Binghamton. 1991.

[VANDERLAAN99] van der Laan, Hans: *De Werkgroep IRT*. To be published. 1999.

[WEISS58] 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.* ACM: New York. 1958.

[WELDON75] Weldon, Jay-Louise: *Implementation strategies for the census data base.* In: [KERR75] p. 589–590. 1975.