Current Research Information as Part of Digital Libraries and the Heterogeneity Problem Integrated searches in the context of databases with different content analyses

Jürgen Krause (Keynote Speaker)
University of Koblenz-Landau and Social Science Information Centre (IZ Bonn), Germany

Abstract

Users of scientific information are now faced with a highly decentralized, heterogeneous document base with varied content analysis methods. Traditional providers of information such as libraries or information centers have been increasingly joined by scientists themselves, who are developing independent services of varying scope, relevance and type of development in the WWW. Theoretically, groups that have gathered current research information (CRI), literature or factual information on specialized subjects can emerge anywhere in the world. One consequence of this is the presence of various inconsistencies:

- Relevant, quality-controlled data can be found right next to irrelevant and perhaps demonstrably erroneous data.
- In a system of this kind, descriptor A can assume the most disparate meanings. Even in the narrower context of specialized information, descriptor A, which has been extracted in an intellectually and qualitatively correct manner, and with much care and attention, from a highly relevant document, is not to be compared with a term A that has been provided by automatic indexing in some peripheral area.

Thus, the main problem to be solved is as follows: users must be supplied with heterogeneous data from different sources, modalities and content analysis processes via a visual user interface without inconsistencies in content analysis, for example, seriously impairing the quality of the search results. A scientist who, for example, is looking for social science information on subject X does not first want to search the social science literature database SOLIS and the current research database FORIS, and then the library catalogues of the special compilation area of social sciences at the library catalogues and in the WWW – each time using different search strategies. He wants to phrase his search query only once in the terminology to which he is accustomed without dealing with the remaining problems.

Closer analysis of this problems shows that narrow technological concepts, even if they are undoubtedly necessary, are not sufficient on their own. They must be supplemented by new conceptual considerations relating to the treatment of breaks in consistency between the different processes of content analysis. Acceptable solutions are only obtained when both aspects are combined. The IZ research group (Bonn, Germany) is working on this aspect in four different projects: Carmen, ViBSoz, ELVIRA and the ETB project. Initial solutions for transfer modules are available now and will be discussed.

Keywords:
virtual library, current research information, content analysis, metadata, heterogeneity, text-fact integration, multimodality, ELVIRA, CARMEN, ViBSoz, ETB

1 Current research information as part of digital library integration

The objective of a specialized virtual library is to enable users to gain integrated access, within a complete system and irrespective of the location and time, to all relevant information in their spe-

cial scientific field – from metadata at the individual document level through to full text which can be called up online. Currently existing media breaks in the acquisition of literature and in searching must be overcome. An integrated user interface has to ensure user-friendliness and, compared with existing solutions, makes it easier for the users to get information of different data types and sources.

Digital libraries combine scientific information from traditional libraries, information and documentation centres with their specialized databases, and from the WWW. They are hybrid libraries containing electronic and printed texts, as well as information with completely different modalities and medialities which, in addition to literature, include current research information, factual databases, photos, graphics, films and teaching materials. They therefore combine rules and standards which are each valid for the different worlds that have to be integrated.

Current research data are today treated like text data and are therefore accessed by descriptors. However, they are rather a mixed form consisting of factual as well as textual information. From the aspect of digital libraries, they are one subset which must interact with all other subsets. Users looking for text documents concerning a special topic might want to retrieve all the data available in different databases. They might be looking for time series, survey results, current research data or a list of experts. Therefore, future innovative research information systems should be embedded in the context of extensive integration of different data types, modalities and medialities. The concept describing this form of integration used here is the digital library concept.

The CRI problem of integration belongs to the same class of problems as that of integrating texts combined with different indexing methods. At the first level a CRI system will face documents which were made accessible through various different methods of content analysis. The use of detailed thesaurus for special purpose collections are one example. At the second level the differences between CRI content analysis and those of literature databases (and others) have to be handled.

1.1 Libraries and specialized information centers

Libraries use the traditional method of standardization to provide users with integrated access to their collections. In Germany alone, there are a large number of relevant standards for libraries. In addition to DIN and ISO standards, there are library regulations and exchange formats based on these standards[1]. With regard to bibliographical description (non-subject indexing), DIN 1505 “Titles of literature” form the basis for the “Alphabetic Cataloging Rules, Scientific Libraries (RAK-WB)”. The related German exchange format is the MAB format. In spite of this implemented standardization, the results of non-subject indexing are often heterogeneous. If German collections are also offered together with Anglo-American collections, this produces the problem of comparison between RAK-WB and AACR2 (“Anglo-American Cataloging Rules”) and the related exchange format MARC21 (Machine Readable Cataloging Records). The existing conversion software provides less satisfactory results, which is why libraries often do not convert the already recorded documents in AACR2 and MARC21, and instead re-record them.

Differences in bibliographical descriptions were not so pronounced and of little consequence in using printed catalogues or index cards. During an integrated search in digital libraries, however, these types of heterogeneity already result in the loss of relevant documents.

Much more critical is the situation with respect to content analysis by means of classifications or thesauri (verbal content analysis). In this case Germany has no longer been able to implement uniform standards. Although the RSW rules (“Keyword Catalogue Rules”), which were used to create a standard keyword file, the SWD, have become widely accepted in subject indexing dur-

1 The following comments are based on Gömpel/Niggemann 2002, Krause/Niggemann/Schwänzl 2002, Geisselmann 1999

ing the past years, special libraries and specialized information systems normally use their in-house-developed special thesauri (the social sciences, for example, with the SOLIS thesaurus). The changeover, for example from SWD or the SOLIS thesaurus, which is also available in English, to the world's most widespread American "Library of Congress Subject Headings" (LCSH) has not taken place and is also not trivial.

German scientific libraries have been unable to agree on a system for classifications. Although the Regensburg composite classification, the GHB listing system and the basis classification are all fairly widespread, it is difficult to convert them to the world's most commonly used system, i. e. the "Dewey Decimal Classification" (DDC), or to the Library of Congress Classification.

Again, specialized libraries and information centres frequently use their own individual special classifications.

The contents of a large part of library collections have also not even been indexed. Geisselmann 1999: 46 mentions a quota of 40 % to 60 %. According to Krause/Niggemann/Schwänzl 2002, the percentage of verbal subject indexing ranges between around 12 % in the south-west German library network and around 46 % in the Bavarian library network. This means that only the terms of the title, and possibly the full text, are available for searching.

Heterogeneity of this kind is also typical for the data of specialized information centres. Since they only handle a few specialized disciplines, different rules are required for more extensive subject indexing than for general libraries. Every discipline has in turn its own rules which are not matched to associated disciplines, a situation that leads to unavoidable interoperability problems in intersecting areas.

The same is true of CRI, also if connected with specialized libraries and information centres. To give some examples: CERIF, the Common European Research Information Format, developed by members of EUROCRIS, recommends for subject indexing the „ORTELIUS" thesaurus. The Information Centre for Social Science Research IZ in Bonn (Germany) uses the SOLIS thesaurus and the SOLIS classification (both German, English, French and Russian) for the CRI system FORIS (Germany, Austria and Switzerland), which are also used for the literature database SOLIS. CRI data at the WWW homepages of the universities in Germany have no standardized thesaurus or classification at all. With respect to classification in Austria most accepted is „OESTAT" which is based on the „Fields of Science and Technology" of the UNESCO (proposed in the OECD Frascati Manual 1980; http://www.statistik.at/). This classification is used in the Austrian database AURIS which contains all university projects of the country (www.auris.ac.at).

The Belgium CRI database IWETO: does not have a thesaurus or authority list for keywords at all (like the Danish National Research Database). The keywords are freely chosen by the researchers when providing the information. As a subject classification IWETO uses a specific version of the CERIF-disciplines and an older version of the NABS.

It is interesting that at the end of 2001 German libraries decided to at least partially eliminate the heterogeneity between German and Anglo-American rules by trying to replace RAK-WB by AACR2 ("Anglo-American Cataloging Rules") and the associated exchange format MAB by MARC21 (Machine Readable Cataloging Records) (see Gömpel/Niggemann 2002 for discussion). This decision of the German "Standardization Committee" was approved by the Library Committee of the German Research Association. Thanks to a joint arrangement and agreement, the traditional form of standardization therefore applies once again to rules for an important area. No such attempts have been made with regard to interoperability between libraries and specialized information centers.

1.2 WWW

For many years, CRI and every type of textual documents have been found to an increasing extent in the Internet websites of university institutes and research institutions. This information

will also be accessed in a digital library. The Internet is therefore extending the previously described system of a wide range of different standards for scientific information by adding an increasing number of new formats[2]. Uniform Resources Identifiers are used as a common organization system, but they regulate little more than the use of characters. Access mechanisms to WWW objects use standardized protocols such as tcp/ip, http, ftp, telnet, mail, news, etc. Subject indexing is not one of the standardization objectives of these techniques – bibliographical description in the conventional sense does not exist.

Parallels to the bibliographical description and subject indexing of information and documentation centres and libraries are found in the WWW under the term "metadata" (as a starting-point for knowledge representation techniques through to the "semantic web"). The DublinCore Meta Data Initiative (DCMI), which probably has the broadest basis in vocabulary development, pursues a similar strategy as the HTML standard with slogans such as "everything optional, everything repeatable". The HTML standard contained the following statement: "WWW parsers should ignore tags which they do not understand, and ignore attributes of tags they do not understand."

This basic attitude, which is not found in the traditional standardization of libraries and specialized information centres, stems from the fact that the WWW has hardly any means of bringing pressure to bear to ensure that standards are implemented. Non-consideration of the proposed rules must therefore also be modeled right from the very beginning. (http://www.w3.org/History/199921103-hypertext/WWW/MarkUp/MarkUp.html).

1.3 General Trends

This review of standardization activities in digital libraries shows:

All attempts, especially in connection with the content analysis of data sets, follow the traditional standardization philosophy which is also closely attached to the theoretical principles of content analysis in information and library sciences. Documents are recorded uniformly based on a standardized, intellectually controlled method, which is developed and implemented by a central organization. In this concept maximum priority is attached to data consistency, so that the user (idealiter) is always faced with a homogenized data set. The widest possible regulation will ensure attainment of the consistency that is regarded as necessary for user questions. In subsets such as library catalogues and specialized databases, this model certainly turned out to be a feasible method which has proved its worth over the past twenty years. However, the general conditions have changed. The technological, economic, political and social changes in recent years have produced trends and opinions which contradicted this models in some aspects.

In spite of all plausibility of the associated advantages, the demand for standardization has only actually been implemented in some subsets. At the latest since the introduction of specialized databases, whose central new content were metadata to newspaper articles, users have had to work with different content analysis concepts. A large number of subject indexing thesauri and classifications for the aspired-to subcomponents in digital libraries up to automatic indexing solutions are now represented in all disciplines.

The discussion in Germany concerning the conversion to American standards does not contradict this general trend. It involves an important subsegment where it remains clear that consistency is increased, but will also not be sufficient in a best case scenario for the interoperability required by virtual libraries.

2 See Krause/Niggemann/Schwänzl 2002: Section 1.3

1.4 Consequences and potential solutions

The above observations harbor serious consequences for digital libraries: Narrow technological concepts, even if they are undoubtedly necessary, are not sufficient on their own (see Krause 1996 for more details). They must be supplemented by new conceptual considerations relating to the treatment of breaks in consistency between the different processes of content analysis. Acceptable solutions are only obtained when both aspects are combined.

The existing extensive heterogeneous databases have been developed in very different ways. Examples of the now attained technical integration of heterogeneous databases in Germany[3] are the KOBV library network (see Lügger 2000 and http://www.kobv.de/se/cont.html), the virtual library network in North Rhine-Westphalia or the virtual library catalogue DVK (http://www.ifs.tu-darmstadt.de/dvk.html[4]).

What the KOBV and comparable projects have not been able to do so far is take sufficient account of the different content analysis processes of the document subsets. The existing conceptual gap mainly concerns the differences in meaning between the distributed resources tied together in a virtual library:

The wide range of breaks in consistency are shown very clearly, especially through the technological merger at several points of libraries, information and documentation centers and WWW sources. A descriptor A may assume the widest possible range of meanings in such a system. In the narrow field of specialized information a descriptor A, which was determined with a great deal of expert intellectual effort from a highly relevant document set, cannot, for example, be equated with term A which provides automatic indexing from a marginal area.

Nowadays, hardly anyone still believes that the document area of digital libraries could be homogenized through global standardization agreements across all subsets, reduced again in organizational terms to a few players or designed by means of an hierarchically organized cooperation model. On the contrary, current concepts start from even greater decentralization in the creation, bibliographical description, subject indexing and distribution of documents, which means that “anarchistic tendencies” will increase still further.

Standardization attempts such as the connection of German-speaking countries to AACR2 are an important step towards provider-overlapping search processes in the heterogeneous data area. However, they do not produce any continuous homogeneity of data, they improve them to some extent. The remaining and unavoidable heterogeneity must therefore be countered by means of different strategies. New problem solutions and further developments are therefore necessary in two areas:

- Metadata
- and the methods of dealing with the remaining heterogeneity. Matching transfer modules must be specified between the individual data types (e. g. literature databases and Internet sources). These modules must counter the lack of standardization by means of automatic methods.

The demands in both areas are closely connected. First of all, the lost consistency will be partially created through the continued development of metadata. Secondly, heterogeneity treatment methods will be used to interrelate documents with different levels of data relevance and subject indexing. The general premise formulated in Section 1 therefore applies to digital libraries: from the aspect of the remaining heterogeneity, their standardization attempts are showing the first visible signs of success.

---

3 See also the Stanford Digital Libraries Project http://www-diglib.stanford.edu/diglib/WP/PUBLIC/DOC104.html and Roscheisen et al. 1997.

4 In ViBSoz “Virtual Social Science Library” the DVK is combined with the SOLIS and FORIS databases of the IZ, Bonn, the library catalogues of the University of Cologne and the database of the Friedrich Ebert Foundation. This takes place with consideration of the heterogeneity problems as discussed here.

2 Handling the Remaining Heterogeneity

We can assume now that standardization efforts such as Dublin Core (DC) or the change from the German RAK standard to AACR2 (see Chapter 1.1) are a useful precondition for comprehensive search processes in the heterogeneous data pool, but despite voluntary consultation by everyone participating in the information process, there is still no sufficient homogeneity in the creation of documents and never will be. The remaining and unavoidable heterogeneity must therefore be treated adequately. The IZ group is working on this aspect in Carmen[5]. It is also one central theme of other current projects of the research department of the IZ: ViBSoz[6], and ELVIRA[7].

The model outlined below represents a general framework in which certain categories of documents with different content analysis are analyzed and referred to one another. The central features are transfer components between the different forms of content analysis, which take account of semantic-pragmatic differences. They interpret the integration between the individual document sets with different content analysis processes (including automatic indexing) on a conceptual basis by cross-referencing the conceptual world of specialist and general thesauri, classifications, etc. The system must know, for example, what it means when term A was used from a specialist classification or a thesaurus for intellectual indexing of a magazine article, but the WWW source could only be automatically indexed. The classification term A could probably only be found by accident in the article, and yet there are conceptual references between both which have to be evaluated.

It is therefore necessary to develop transfer modules between data sets. These modules will permit transfer both in technical and conceptual terms. In these considerations it does not matter, in principle, whether the semantic-pragmatic differences between two texts, between a text and current research information, between a text and the results of a survey in tabular form or a video archive have to be bridged or alternatively between multilingual sources.

In principle, there are three methods which have to be checked and implemented in relation to their effectiveness in individual cases. None of the methods is solely responsible for transfer. They are interlinked and interact with one another.

- The development of an intellectually gained crosswalk of various indexing and classification systems ("cross-concordances").
  Cross-concordances have already been used and implemented in the current projects of IZ and will not be described here (see Krause/Plümer/Schwänzl 2000).
- Statistical methods and neural networks.
- Deductive methods. They can be seen in parallel with artificial intelligence methods in expert systems; and will also not be discussed here.

Using cross-concordances, a simple rule transforms one term into another term suitable for mapping. Although this method works well for some terms, empirical studies have shown that classifications and thesauri are often so different that such simple mappings will be restricted to a sub-

5 CARMEN investigates the conceptual problem of heterogeneous data stores on the basis of an exemplary data pool from the contents of major publishers' servers linked to electronic information primarily in the fields of mathematics, physics and social sciences (see Krause/Schwänzl/Plümer 2000).

6 The "Virtual Social Science Library Project" concentrates on the connection of different library catalogues with the documents of the SOLIS literature database (see Kluck et al. 2000). It is part of the digital library research program of the German Research Association (DFG).

7 "Elektronisches Verbandsinformations- Recherche- und Analyse-System = Electronic Retrieval and Analysis System for Industry Associations" (funded by the German Federal Ministry of Economic Affairs). ELVIRA started as an online system for time series used by major German industry associations to provide information to their member companies. Text retrieval functionality was added in 1998 to search text collections mainly relating to foreign trade and to address fact–text integration. The system is now used by companies in more than 350 installations (see Krause/Stempfhuber 2001).

set of the terms in question. Therefore, a second strategy was integrated in combination with the first transformations which rely on vague methods already successfully applied in information retrieval (IR).

In ELVIRA the texts are indexed automatically using a probabilistic standard algorithm from FULCRUM. Contrary to this, the time series are indexed intellectually using a hierarchical nomenclature (fact thesaurus). The statistical-based crosswalk has to transform these meaning differences of the different content analysis procedures of textual and factual data. In ViBSoz the statistical crosswalk is used between two different thesauri (SWD as a universal thesaurus and SOLIS as a special thesaurus). Unlike cross-concordances, the statistical transformation is not based on the general semantics of the intellectually acquired term-to-term link. Instead, words are transformed into a weighted vector of terms representing the use of the term in the document pool.

In the following the treatment of heterogeneity by means of transfer modules will be described in more detail.

2.1 Treatment of vagueness in information retrieval

Vagueness in IR is normally countered by using statistical approaches in which vagueness is modeled between the user query and the document set, whereby the document level is regarded as the uniform modeling basis. This is shown most clearly when all documents were automatically indexed. Even if additionally intellectually determined descriptors of a controlled vocabulary were produced, modeling follows this homogeneity demand in principle. The user can either carry out his/her search using one of the two term groups and then base his/her search strategy on the controlled vocabulary or alternatively via the free text terms of automatic indexing using another strategy. If he/she chooses to perform a search via both term areas, differentiation in the match is not distinguished, i. e. it is treated as if the semantic differences in both content analysis methods do not exist.

The problems become even clearer when we use digital libraries and link, for example, a social science literature database such as SOLIS with its own controlled vocabulary (IZ thesaurus and IZ classification) with library catalogues, for instance in the ViBSoz Project with the documents of Cologne University Library (USB Thesaurus), which is being developed intellectually according to the keyword list of the German Library (DDB) in Frankfurt. A comparison of two such thesauri or classifications reveals that vagueness already occurs at the semantic description level of two document sets to be integrated in the search, and not just in the user query. Terms in a library classification with its controlled vocabulary form a separate description level. It cannot simply be translated on a 1:1 basis into the terms of another classification, e. g. from the area of specialist information systems. The meaning of a descriptor A in the library classification is different from the meaning of the same term in another classification or even in a thesaurus such as the IZ Thesaurus of the SOLIS social science literature database, even if a simply “vague” connection exists. These vagueness relations can be integrated in a non-differentiated way into the modeling of the IR process. In this case vagueness between the user query and documents will be modeled without explicitly treating differences at the document level between two heterogeneously developed document sets separately by means of transformation modules. Therefore, we speak here about a “one-step” method.

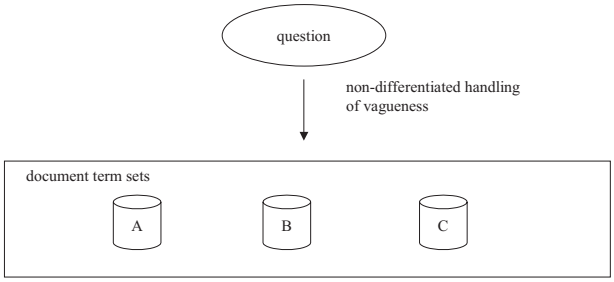

Fig. 1: One-step method

2.2 "Two-step" method and transfer modules

An alternative "two-step" method plays an important role in the projects of the IZ (ELVIRA, ViBSoz and CARMEN). It is based on the thesis that heterogeneous document sets should first be interlinked through transfer modules (vagueness modeling at the document level) before they are integrated in the superordinate process of vagueness treatment between documents and the query (the traditional IR problem). If, for example, three heterogeneous document sets have to be integrated, transfer modules bilaterally treat the vagueness between the different content analysis methods. The aim behind this form of vagueness treatment, which differs considerably from the procedure used traditionally in IR, is to produce greater flexibility and target accuracy of the overall procedure through separation of the vagueness problem. Different forms of vagueness do not flow uncontrolled into one another, but can be treated close to the causal interface (e. g. the differences between two different thesauri). This firstly appears more plausible in cognitive terms and secondly permits the combination of a wide range of modules for treatment of vagueness. Together, these modules can become effective during retrieval using heterogeneous data sets.

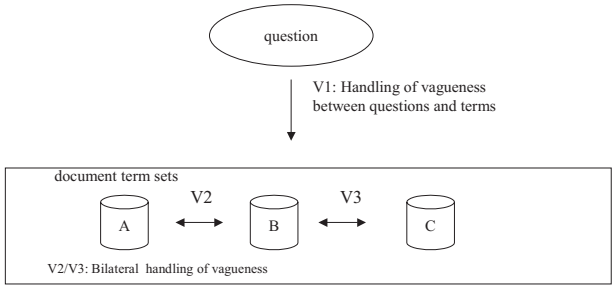

Fig. 2: Two-step method

This procedure appears to be highly promising, especially in view of the broad background of empirical knowledge that IR processes differ more in the volume of results than in the quality of the evaluation parameters such as recall and precision. For example, a probabilistic method can be combined with neural transformation modules in the match between the user query and documents. Alternatively, neural transformation modules can also be combined with IR processes based on Boolean algebra. But even if the transformation modules between the heterogeneous

document sets use the same similarity function as at the IR level between the user query and documents, the results between the "one-step" and "two-step" methods will probably differ.

2.3 Integration of the results of vagueness treatment through transfer modules

The architectures which were introduced as the "one-step" method and which model vagueness in an unspecified manner between query terms and document terms have no integration problems whatsoever since the query terms occupy the input layer of the neural network and the documents the target structure. However, vagueness relations occur in several places in the two-step method. With just two heterogeneous document sets, it is necessary to answer the question of how the vagueness treatment of "query $\rightarrow$ document set" can be combined with that between document sets.

2.3.1 ELVIRA: directed transfer

ELVIRA permits access to data on production, foreign trade, the economy and economic structures. The time-series fact tables are not addressed via their cell values and the table names, but indirectly via intellectually assigned descriptors relating to three categories, i. e. the subject in question (e. g. export), industry/product (e. g. microwave ovens) and the country. User tests quickly showed that association customers demand both time series and textual information sources to solve their problems. The transfer problem arises because different indexing methods are used in time series and texts. Time series are indexed intellectually and the descriptors used are assigned hierarchically (e. g. by nomenclatures of the Federal Statistical Office). Some texts are indexed automatically, other ones additionally with the same nomenclatures as the fact tables. In this case not only standardized thesaurus terms, but also words which occur in a text are valid search terms. Texts also appear to refer less to individual products, as is the case, for example, with time series of deeply structured production statistics. Texts can often only be found at higher aggregation stages (e. g. for drive systems or electric motors as a whole) and not for special types of motor such as direct-current motors.

In the case of text-fact integration like in ELVIRA, simple directed transfer appears to be the rule. The user first looks for facts and then wants to obtain evidence of the associated texts (and vice versa) in an iterative search step (or right from the beginning). In a "two-step" architecture simple directed transfer does not place any special demands on the integration of both vagueness treatment modules. However, the difference in the function of both architectural concepts becomes clear in this case. Traditional term extension strategies therefore resemble the transfer modules discussed here. But since they work within the meaning of one-step methods, they have different impacts, even if the same mathematical processes were used. This becomes even clearer and - in the impacts - more serious when several heterogeneous text sets are integrated.

2.3.2 Transfer between heterogeneous text sets

The parallel with 2.3.1 is produced in text retrieval when, for example, an IZ user, who has precise knowledge of the IZ Thesaurus and gears his/her search strategy according to this thesaurus, also wants to obtain the texts of Cologne University Library without knowing anything about the thesaurus used there. This technique was implemented as one of the first transfer modules of ViBSoz. However, this will probably be the exception rather than the rule.

More than two document sets with different indexing methods can firstly be expected in heterogeneous text searches. Users can also be expected, who are not incorporated in any of the used content analysis methods in such a way that they can direct their search strategies according to a special standard.

This does not matter in the one-step method as no distinction is made between bilateral transfers. However, the two-step method must ensure that the advantages of measuring vagueness differentiated between two indexing systems are not lost as a result of the way in which these values are integrated in the internal search. The simple method of non-specific extension of the query term is therefore no longer sufficient. If the query is supplemented, for example, by a term from the vagueness definition B→ C, this term will only become effective in C, but not in A, for instance. Instead of a globally effective term extension strategy, this makes it necessary to extend the query terms on a differentiated basis and operate on different subquantities of the heterogeneous document sets.

3 The consistency problems of digital libraries as a general standardization problem

A general discussion concerning the limits of current standardization problems in industry and administration[8] is being presided over by the responsible national authorities (DIN in Germany), European bodies (EUN) and international organizations (ISO). This discussion shows that the problems described here for digital libraries are of a very general nature. However, there are clear indications that the traditional methods of standardization are coming up against limiting factors in an increasing number of areas (not only in digital libraries) – in spite of all the detailed improvements in the methods. On the one hand they appear indispensable and substantially improve quality and cost-effectiveness in subsets. On the other hand they can still only be partially implemented, with increasing costs, within the framework of global provider structures and changed general conditions. The current standardization concepts must therefore be revised. The remaining and unavoidable heterogeneity must be countered by different intellectual and automatic methods of retrospective conceptual integration. A new way of thinking regarding the existing remaining demand for consistency retention and interoperability is necessary. It can be described by means of the following premise: Standardization must be considered from the aspect of the remaining heterogeneity. A solution strategy, which takes account of general current technical, political and social conditions, can only be obtained through joint interaction between intellectual and automatic methods relating to heterogeneity treatment and standardization. In specific terms, this means the following for standardization projects: If standardizations cannot be implemented or can only be implemented to a partial extent in subsets in a reasonable amount of time, every remaining detail must be analyzed specifically to determine the consequences of the lack of standardization and how the remaining heterogeneity can be at least roughly countered by means of automatic or intellectual methods. Any resulting costs and losses of quality must be compared with the time expenditure and success prospects of other intensive attempts to attain standardization.

4 Literature

Geißelmann, Friedrich (1999): Zur dritten Auflage der RSWK, Bibliotheksdienst, 33. Jg., H. 1

Gömpel, Renate, Niggemann, Elisabeth (2002): RAK und MAB oder AACR und MARC? Strategische Überlegungen zu einer aktuellen Diskussion, ZfBB 49.1.

Kluck, Michael; Krause, Jürgen; Müller, Matthias; in Kooperation mit Schmiede, R.; Wenzel, H.; Winkler, S.; Meier, W. (2000): Virtuelle Fachbibliothek Sozialwissenschaften: IZ-Arbeitsbericht 19, IZ Sozialwissenschaften, Bonn.
http://www.bonn.iz-soz.de/publications/series/working-papers/index.htm#Virtuelle

Krause, Jürgen (1996): Informationserschließung und -bereitstellung zwischen Deregulation, Kommerzialisierung und weltweiter Vernetzung. (IZ-Arbeitsbericht Nr. 6) Bonn 1996.

8 to appear in http://www.din.de

Krause, Jürgen; Niggemann, Elisabeth; Schwänzl, Roland (2002): DIN-SICT Papier „Strategie für die Standardisierung der Informations- und Kommunikationstechnik (ICT)“ (to appear)

Krause, Jürgen; Plümer, Judith; Schwänzl, Roland (2000): Content Analysis, Retrieval and Metadata: Effective Networking for Mathematics, Physics and Social Sciences, RC33-Session “New Conceptual Developments in Information Systems and the WWW”. Proceedings der Fifth International Conference on Social Science Methodology, October 3 - 6, 2000. Köln (CD-ROM).

Lügger, Joachim (2000): Über Suchmaschinen, Verbünde und die Integration von Informationsangeboten, Teil 1: KOBV-Suchmaschinen und Math-Net. ABI-Technik 20, Nr. 2, S. 132 - 156.

Roscheisen, Martin; Baldonado, Michelle; Chang, Kevin; Gravano, Luis; Ketchpel, Steven; Paepcke, Andreas (1997): The Stanford Infobus and its Service Layers. Augmenting the Internet with Higher-Level Information Management Protocols
http://www-diglib.stanford.edu/cgi-bin/WP/get/SIDL-WP-1997-0065
[eingesehen: 26.05.98]

5 Contact Information

Jürgen Krause
Social Science Information Centre (IZ Bonn)
Lennéstr. 30
D-53113 Bonn
Germany

e-mail: krause@bonn.iz-soz.de
http://www.uni-koblenz.de/~krause/