DOCUMENT RESUME

ED 450 762 IR 058 018

AUTHOR	Geisselmann, Friedrich
TITLE	The Indexing of Electronic Publications--Ways out of Heterogeneity.
PUB DATE	2000-08-00
NOTE	8p.; In: IFLA Council and General Conference: Conference Proceedings (66th, Jerusalem, Israel, August 13-18, 2000); see IR 057 981.
AVAILABLE FROM	For full text: http://www.ifla.org/IV/ifla66/papers/173-181e.htm.
PUB TYPE	Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	Academic Libraries; Access to Information; *Classification; Decentralization; *Electronic Libraries; *Electronic Publishing; Foreign Countries; *Indexing; *Information Services; Metadata; Models; *Thesauri
IDENTIFIERS	Concordance (Data); *Electronic Resources; Germany

ABSTRACT

This paper begins with some general remarks about the indexing of electronic publications. Characteristics of today's information world are highlighted, including decentralization, distributed data collections, and heterogeneous data structures and indexing procedures. Three possible models for managing access to information are summarized: a centralistic organization; a network of scientists; and standardization, acceptance, and dissemination of metadata. The shell model for indexing electronic resources is presented. The CARMEN project of the Universitatsbibliothek Regensburg (Germany), which deals with the indexing of digital publications, is then described, focusing on cross concordances between different classifications and thesauri. Similar links between a universal thesaurus and specialized thesauri are also discussed. (MES)

IFLANET Search Contacts
International Federation of Library Associations and Institutions
Annual Conference

66th IFLA Council and General Conference

Jerusalem, Israel, 13-18 August

Conference Proceedings

Code Number: 173-181(WS)-E
Division Number: IV
Professional Group: Classification and Indexing: Workshop
Joint Meeting with: -
Meeting Number: 181
Simultaneous Interpretation: *No*

The indexing of electronic publications - Ways out of heterogeneity

Friedrich Geisselmann
Universitätsbibliothek Regensburg, Regensburg Germany

Paper

I would like to make in my paper, first of all, some general remarks and deal then with a concrete project, CARMEN, in which our library is involved.

Today's information world is characterized by strong decentralization, distributed data collections and heterogeneous data structures and indexing procedures.

Traditional producers and brokers of scientific information are scientists, the publishing houses, the libraries and information service points. Their functions remained unchanged in the main until some years ago. What are changing dramatically today and require significant advancements and adjustments are the basic conditions. The present situation is shaped by a deep change in the entire fundamental basis of information technology, which not only forces technological adjustments, but also opens up new forms of information services and suggests other than the traditional forms of interaction between scientists, publishing houses, libraries and information service points.

Besides the traditional information providers -- publishers with their print media, libraries with their holdings indexed according to intellectually assigned classification schemes and specialist information centres which offer their databases via hosts -- scientists themselves are also strongly represented and have been developing in all these areas via the WWW independent services with the most varied coverage, relevance and means of indexing. Generally speaking, groups can occur anywhere in the world, which collect information on specialist fields. This is

regarded by some as progress : the traditional bottlenecks of the publishing houses, libraries and specialist information centres are eliminated ; information flows unhindered from scientist to scientist. There is however also the contrary opinion : a consequence of this revolution are different consistency breaches (Figure 1):

- Relevant, quality-controlled data stand next to irrelevant and possibly demonstrably false ones. No expert system ensures a differentiation between ballast and potentially desired information.
- Descriptor A in one such system can assume the most varied meanings. Not even in a narrow field of specialist information can a Descriptor A, with a highly relevant document collection which has been intellectually and qualitatively determined with great effort, be equated to the term A, which delivered automated indexing from a peripheral field.
- Quite different documents with comparable intensity are offered in one and the same search result : the announcement of a lecture on algebra is beside a textbook on algebra and beside a special treatment from this field. In former times these media were separated by different forms of indexing.

Despite such problems, the user will want to access the different data collections, irrespective of what method they have been indexed by or in which system they are offered. Even in the world of decentralized, non-homogenous data collections, the user will rightly demand of information science that he is ensured the receipt as far as possible of only the relevant documents and if possible all the relevant ones, which correspond to his need for information. (Figure 2 shows such a system with combined search over library catalogs and special data bases)

How can this be managed? On this question, policy and science give at present three responses, which are to be discussed by example of what the specialist information centres offer.

A model solution envisages a centralistic organization : Due to technical developments, libraries and specialist information centres were 20-25 years ago of necessity centralistically organized and as a consequence of this also conceptually aligned to the centralistic approach to contents indexing. A centrally set-up large computer administered the data. The clientele was served via terminals or off-line via inquiries to a central point. The theoretical basis of the contents indexing corresponded to this. In a standardized, intellectually controlled procedure, which the central point developed and carried through, a uniform creation of the documents took place.

A second model is that of a network of scientists : in 1995 some learned societies in Germany united to form an "IuK-Kommission" (Commission for Information and Communication). In the cooperation agreement it is stated : "There is a concensus in the view that new IuK structures within the academic field should be organized on a 'distributive basis'. Information in the future will not exclusively be supplied by the traditional information providers, but increasingly by the creators of the information themselves, i.e. by the scientists in the different subject areas, specialist fields, institutes and learned societies. The present, primarily centrally organized supply of electronic information should be supplemented and enriched by decentralized information. The learned societies aim in their planning and development of distributive information structures at the creation of an as complete as possible, structurally clear and economic supply of information in their sciences."

A third model aims at the standardization and acceptance and spreading of metadata. These are a prerequisite for provider-overlapping search processes in a daily ever-increasing decentralized information world. They try to re-establish in part the missing data-homogeneity by voluntary agreements with all those

participating in the information process.

However, this can succeed only partly. However successful the introduction of metadata will be in a subject area, the remaining heterogeneity of the different types of the contents indexing (automatic, different thesauri, different classifications, differences in the categories created) will be too great to overlook. In all but a few exceptions today, however, exactly this occurs. Descriptors, which were determined in the most different circumstances of contents indexing and have therefore a different meaning and relevance in each case, become directly connected by search engines via (technically) distributed databases, which is a reason for the unsatisfactory results of today's system implementations.

The attempt to reconstruct the consistency of indexing by librarians or documentation centres (e.g. in CORC) is possible only for parts of collections. The question arises whether the request "Let us catalogue the Internet" is correct, and whether it is at all realizable. Doesn't the cataloguing of electronic resources have to be limited to particularly important documents? (Figure 3) Over and above these documents do we not have to use different tools?

As a contrasting model, the shell model was developed by Jürgen Krause[1] (Figure 4) In general, each deregulation without coordination control leads to anarchic structures. This applies also to contents indexing. The more strongly deregulated it is on all levels, in order to get away from centralistic structures, the more important the authority control is. The disintegrating units and newly added groups must be referred flexibly one to the other, without the requirement for control and standardization of the former model conceptions being revived again under new terminology. Such a model of information indexing allows for different levels of data relevance and contents indexing, which are referred one to the other in a common information system by transfer components. Standardization and qualitative requirements are not centralistically implemented, but coordinated and administered.

- The innermost shell contains the core of the highly-relevant documents. It is indexed as deeply and at the highest level of quality as possible. Quality control is located in the hands of the coordinating information service point. Neither for scientific nor for organizational reasons can this shell be dispensed with. Only the resulting data consistency of a core area creates the incentive for further partners to use the following shells.
- The second shell loosens the relevance conditions and in parallel to this the requirements of the quality of the contents indexing. Shell 2 could, for example, contain documents of the libraries which are indexed according to a standardized thesaurus, but offered without abstracts.
- Shell 3 could contain all documents, whose relevance is lower in comparison to shells 1 and 2 (e.g. peripheral areas or missing checks for consistency) and/or indexed according to other standards (e.g. different thesaurus).
- Shell 4 would contain unchecked self-announcements in the WWW.

Compared to the ideal model of continuous consistent indexing, data consistency and thus the quality of search sink even under the best conditions. Compared with a realistic scenario the abundance of data thus attained balances out the consistency breaks.

This is where the CARMEN project (http://www.mathematik.uni-osnabrueck.de/projects/carmen/) comes in. This project is financed within the framework of the German program for digital libraries Global Info. CARMEN, in particular, deals with the questions of the indexing of digital publications.

It is necessary on the one hand to strengthen the standardization within the field of metadata, and on the other hand to take into account and reconcile the further existing heterogeneity in the retrieval. But both machine procedures (quantitative-statistical and deductive approaches) as well as intellectually compiled Crosswalks between different classifications and thesauri are necessary.

I will deal first with the last point. On this work package our library, the Specialist Information Centre for Social Sciences in Bonn and Die Deutsche Bibliothek (German National Library) in Frankfurt are cooperating.

[1]Jürgen Krause: Polyzentrische Informationsversorgung in einer dezentralisierten Informationswelt. In: Nachrichten für Dokumentation 1998, S. 345-351.

The starting point is this : in libraries and specialist information systems different classifications and thesauri are used. Thus inter-disciplinary and multiple-database searching is made much more difficult. The user, who, for example, first searches in a library catalogue in Regensburg, afterwards in one from Lower Saxony or the USA and subsequently articles from the literature database of an information service point, must operate in each case with differing search terms and varying search logic, and thus an efficient search is hardly possible. Usually the user knows only the classification or the thesaurus, with which he primarily operates. This problem is intensified, if the different library catalogues and literature databases are connected technologically in such a way that the user can access them with a uniform search screen. This applies also to the use of different indexing systems in metadata.

The aim is to enable an integrated search from a subject standpoint in distributed data collections. Thereby the conceptual differences in thesauri and classifications used must be borne by cross concordances. This requires:

- the investigation of the methodology of cross concordances between classifications or thesauri.
- the programming of a procedure, how such cross concordances between different classifications or thesauri available in the Internet can be depicted.
- the development in prototype of such cross concordances for certain subject areas and selected classifications or thesauri.

A parallel methodology with classifications and thesauri enables the gaining of knowledge of the common or diverse problems, and methods for finding solutions to the different indexing procedures and subject areas. This guarantees the prototype character of the investigation. The solutions should be able also to be applied to other classifications and thesauri not included in our investigation.

The cross concordances refer to classifications/thesauri, which represent closed systems in themselves. Between different classifications/thesauri it must be possible to navigate with the help of the cross concordance. It is a fundamental fact that the classifications/thesauri are driven by different institutions and not by the institution which operates the concordance. Access must thus also be possible via the Web to other computers. This simplifies also the problems of updates, an important problem, if one regards the permanent operation of such a system. However, this presupposes that the systems are available on the Web. The data structure is represented in Figure 5.

What is significant for the data structure is that the concordance is always formed between two classifications and not all classifications on the one used, or a new hyperclassification. The linking is done via the notation. The linking has thereby

according to our present conceptions a direction, i.e. it is conceivable that the linking in the opposite direction can be created differently.

We proceed thereby in such a way that the hierarchy in the classification is used in order to simplify the concordance. The linking takes place only with coordinate terms. If there are narrower terms in one of the classifications, then these terms are also contained under normal conditions on the opposite side. Thus one only needs to follow the route upwards to a linked term. However, even if the other classification also has narrower concepts, then one can naturally proceed differently. (Figure 6)

Apart from the actual linking, the type of the relationship must be created between the linked notation/descriptors assigned; (Figure 7)

- 1 : 1 relationship (synonymous terms, parallel notation);
- Broader term : narrower term (broader : narrower);
- Narrower term : broader term (narrower : broader);
- Related terms;

A further point, which must be created with each relationship, is the relevance of the linking. This is a roughly estimated value. We divide it therefore only into 3 levels :

- low
- middle
- high

One can determine Precision and Recall empirically or attempt to estimate them.

The methods of a concordance between general classifications and special classifications should be compiled by way of example. In particular, the subjects mathematics and physics were selected because these fit into the overall project. As classification schemes the following are particularly well suited : Dewey Decimal Classification (DDC) and the Regensburg Network Classification (RVK) as well as the Mathematical Science Classification (MSC) and the Physics and Astronomy Classification Scheme (PACS). In addition, for the social sciences a concordance will be created between the RVK and a German classification of social sciences.

Thus two general classifications are selected on the one hand, which are particularly important internationally or in Germany (over 100 users). Such a cross concordance is also for our users of special interest, since it makes the search for users of the DDC in our catalogues possible and allows for the carrying out of external work on the classification. The special classifications are in each case in their respective areas generally recognized.

The problems are on the one hand with the subject overlapping of the special classifications (MSC, PACS). The two classifications overlap to a substantial extent also within the core area of the respective subjects. On the other hand, it is very interesting to form a concordance between the strongly specialized and the general classifications (e.g. MSC-DDC). This enables the transfer in queries from the respective special information systems, e.g. the databases of MathSciNet and MATH to a library catalogue, or vice versa, to a certain extent.

So far a tool has been implemented for the intellectual creation of such cross concordances. CarmenX (Figure 8). Features include the

- cooperative handling at distributed locations

- use of a WWW Browser on the client page
- decrease in expenditure for inputting by using text already entered
- possibility of working on different classifications.

So far we have stored the RVK, PACS, MSC and the Classification for the Social Sciences; the DDC not yet, because we still have no data supplied from OCLC.

We operate with the following technology:
WWW server Apache
Relational database system: MySQL
Server-sided script language PHP
Client-sided Frames and Java Script

Subtopic Thesaurus:

Here in a similar way a link will be compiled between a universal thesaurus and several specialist thesauri, concretely between the German authority file of subject headings SWD, the thesaurus for the social sciences, the subject headings material of the DIPF - German Institute for International Educational Research - (developed from the thesaurus for education and the thesaurus for education research).

The problems of thesauri differ from those of the specialist classifications in the

- narrower meaning of the structured figure
- stronger weighting of the selective linking between the individual terms
- greater weighting of the linguistic problems
- greater problems in the ambiguity of terms (homonymy and polysemy)
- the representation of subject heading strings by single terms.

For the subject area of the thesaurus, a different software product SIS-TMS will be used. I do not not wish to go into the details here, since Patrice Landry will be illuminating the topic of thesaurus from a different angle.

A few words on other sections of the project CARMEN, which are treated not by us, but by other partners in the project : The working package of cross concordances presupposes intellectually indexed documents. This is a good starting point, which does not apply by any means universally.

Electronic documents cover many different types of documents : besides scientific texts, descriptions of projects, descriptions of corporate bodies, conferences, tables of contents for periodicals, advertisements, databases with facts, e.g. results of surveys beside texts about these surveys. The documents themselves can be very complex, e.g. with embedded graphs, formulas, programs and links to a multitude of other documents, etc. in the texts. What is important is to make these documents retrievable altogether despite their different structures. A particularly significant problem is presented by databases, which have a quite different structure from textual ones.

Even structurally homogeneous documents are indexed differently. Thus, for example, MPRESS, a world-wide Preprint index for mathematics, contains,

- 40,230 documents
- only 8,927 documents classified with MSC,
- only 22,683 documents with metadata,
- documents without metadata are often stored in layout-oriented formats (PostScript, PDF)

When using meta-tags, varying use must also often be taken into account, details are possible with different formatting.

A significant aim is to design better metadata constructs, and to develop on the other hand procedures to generate metadata automatically.

One can try to attain content and formal categories by means of deductive-heuristic methods from the documents. From this arise many questions, e.g. which Internet pages form, from the point of view of content, self-contained units? How does one retain the context information, i.e. how will the respective home-page be stored?

Between different data collections one can also create relations with quantitative-statistical procedures. These procedures are based on the competition of data in parallel corpora (collections). Neural procedures are also used.

In a further working package, a retrieval system for XML documents is to be constructed. The aim is a system, which can recognize highly-structured documents and a variety of linking structures. This system is to serve as a replacement for Harvest. Such a retrieval system can also process metadata in XML.

Special problems develop thereby in the exchange between such retrieval procedures, which are based on full-text search and database structures. The structure of these procedures varies a great deal.

In summary one can say about these problems: The world of electronic publications is more varied than the world of books. One can network and thereby substantially improve the information. The consequence is, however, that the indexing becomes far more difficult than that for books.

Latest Revision: *October 19, 2000*

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

NOTICE

REPRODUCTION BASIS

☒ This document is covered by a signed "Reproduction Release (Blanket) form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.

☐ This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").