METADATA - A KEY TO DATA ACQUISITION AND INFORMATION RETRIEVAL

Peter J. A. Reusch [1) 2)]

1) Fachhochschule Dortmund - University of Applied Sciences, Dortmund
Peter.Reusch@FH-Dortmund.de
2) IBIES Dortmund-Brьhl PJAReusch@AOL.com

Abstract: *Metadata have a long tradition in several areas. Metadata help to structure data or documents and to classify the contents. Today it is important to recognize the power of existing metadata (thesauri, classifications, ...) and to use such metadata and to expand such metadata for advanced data acquisition and information retrieval. To expand metadata modern learning systems should be used and an open communication model, that integrates different aspects of data and metadata. Results of an information retrieval project are described and a new multi-level model of metadata is introduced.*

Keywords: - *Metadata, Classification, Learning Systems, Information Retrieval*

1. THE HISTORY OF MEATADATA – AND THE SITUATION TODAY

Metadata have a long history in several disciplines. In computer science metadata in a schema definition are used to define databases, to support storage and retrieval of data. In library sciences metadata are introduced to classify books according to content and to support information retrieval. Deweyrs Decimal Classification is a famous contribution in this area. It was published in 1876 and still is a source for new initiatives - for example in the project DESIRE - Development of a European Service for Information on Research and Education. There are several national initiatives to implement new classifications based of Deweyrs approach, for example in the Dutch National Library.

Metadata were used on highly structured data in databases and on weekly structured documents. Here we are only interested in the role metadata can play in content classification and in the support of data acquisition and information retrieval. Here we often meet a struggle.

What is first, the data or the metadata?

The classical librarian will classify any new book according to the predefined classification schema used in his library. The classification process produces “key”-words, that will enable the user to find appropriate books in the library. This classification is time consuming and expensive. Specialists are needed for this process.

Computer scientists often think that such processes are for no real reason today. Information is available in files - if not we can scan papers. We have powerful computers and highly effective software, even based on (artificial) intelligence. This technology can work on “rough” data. The user will find whatever he needs. We donrt need predefined metadata and preclassified documents. We donrt even want such metadata because, preclassification may be wrong. And if a document is not classified properly, you will never find it.

SER Brainware® (www.ser.com) is a learning engine that provides the ability to classify, store, retrieve and extract knowledge. The problem is, when such “brainware”-systems (SER Brainware® and similar systems) operate, the classes that are derived are **not** the classes the human user has in his mind. The results of such “brainware”-systems are not bad, but in a lot of cases there is a gap between “brainware”-classes and “human mind”-classes that makes information retrieval with “brainware”-systems not that easy.

Today we must realize that both aspects are important - classical metadata and metadata derived by learning systems, and that we need both approaches to solve complex problems. “Human” classification can support “brainware”-systems to improve information retrieval, and the information found by learning systems can be used to improve classification, and better classification can again improve intelligent search engines.

In modern information retrieval recall and precision are defined do describe the quality of answers to given requests.

Within a collection of documents we are interested in the set of relevant documents R. The answer of a request for R usually is not identical with R – it’s a set of answers - A -that may include relevant documents.

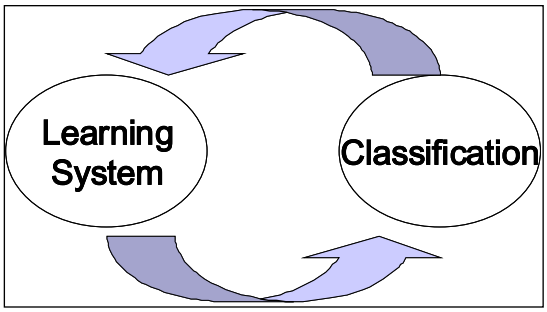

Fig. 1 – ***Metadata improving "Brainware" improving Metadata.***

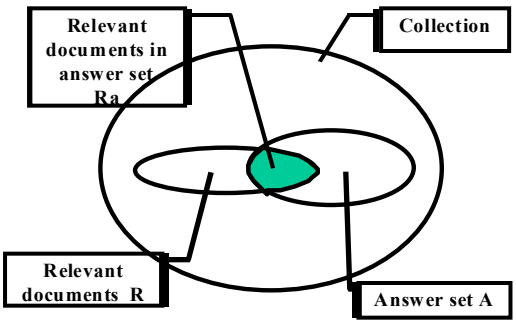

Fig. 2 – ***Reall and Precision.***

Recall is the fraction of relevant documents which has been retrieved

$$Recall = |Ra| / |R|$$

Precision is the fraction of retrieved documents that is relevant

$$Precision = |Ra| / |A|.$$

In optimal cases recall and precision should be close to 1. But in most case we meet to many documents that are not relevant and only a part of the relevant ones.

The question is: What is the reason for bad recall and precision? We must take care of the user and the retrieval system. Good metadata can improve the way the user starts his questions and therefor will improve precision and recall.

2. A METADATA-BASED MODULE SUPPORTING INFORMATION RETRIEVAL

The author took part in a project in the city of Cologne-Germany - where fast and precise retrieval of information from a large database in the city council was needed. In the database there are millions of documents on resolutions passed by the city council. New councilors, administrative staff and even citizens want to known what had been decided in the past. Such users often have no experience in searching for such information. But they have a lot of knowledge about the city, the administration, the political situation and other aspects. And they will look for information according to the knowledge they have in their mind, according to their language, their political background, and the district where they live.

In the past the documents of the city council had been classified - since more than 20 years. Several employees worked of this classification process - and supported inquiries. But these processes had to be improved - regarding costs and efficiency. The author was able to implement a retrieval module based on the established classification and working according to the data driven data reduction method - DDDR. The software module first had been tested by potential users - with good results. Nevertheless before finally implementing the module, a competition was arranged. The users decided that the DDDR-module based on the classification was better than the "Brainware"-system available. The result was a bit surprising. The main reason was, that is much easier to focus on the information you need and get the results with high precision and good recall when you use DDDR-modules. And the key to good precision and recall is, that - at least in the first approach - "human" classes are more convenient for the "human" retriever.

In January 2001 in the city of Cologne we started the first level of the retrieval system based on a special set of documents and a set of about 4000 "human" classes. In the next step these classes will be expanded and integrated within a thesaurus of about 50000 entries.

Such a thesaurus will not only improve the information retrieval process. This thesaurus will support automatical classification of documents. And this thesaurus will be connected with "Brainware"-Systems. "Brainware"-Systems will help to control and expand the thesaurus and DDDR-modules will be able to retrieve more and more information of different sources - in the city of Cologne.

Less than one third of the thesaurus will be specific for Cologne. Most parts can be used in other German cities. The thesaurus will include more words derived from documents of the federal government and similar sources than words only relevant for Cologne - like names of streets and places, etc..

3. LEVELS OF METADATA – A MULTI-LEVEL MODEL

The results of the project mentioned above were stimulating to derive communication models sup-

ported by metadata at different levels. This approach is also very important to support e-Commerce and other businesses.

Assume a given scientific document. It was written by an author, using a special subset of a "natural" language, adapted for his working area, and it was published in a special issue of a journal last year. The document is classified according to the topics selected by the publisher.

A more detailed analysis will show us, that in the first step such a document should have information or links regarding resources, contents and applications.

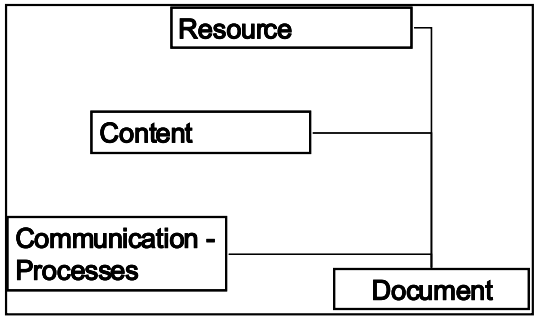

Fig. 3 – ***General Levels of Metadata.***

The W3-Consortium offered a framework for the description of resources, including authors and other elements (www.w3org./TR/rdf-schema).

Regarding the content of a document we must deal with the language of the document in general, with a thesaurus that depends on the area, and finally with the classification.

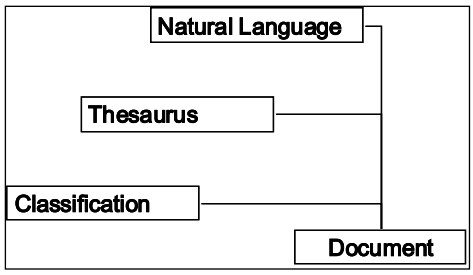

Fig. 4 – ***Content Oriented Levels of Metadata.***

The classification - regarding the descripting words - can be part of the thesaurus. The thesaurus may follow new recommendations - maybe the recommendations of the CERES-project (http://ceres.ca.gov/thesaurus/ RDF.html) with specific words and classes and terms like:

IC	Descriptor within a Category
CAT	Category of the Descriptor
UF	EntryTerm for which the Descriptor is preferred
TT	Topmost Term(s) for the Descriptor
BT	Broader Term(s) for the Descriptor
RT	Related Term for a Descriptor
NT	Narrower Term for the Descriptor
USE	Descriptor that is preferred to the EntryTerm

Mathematical models are available to handle such a thesaurus integrated in a multi-level communication model. Here we can use the results of Rudolf Wille and Bernhard Ganter on algebraic approaches to concept analysis – concept lattices.

Jean-Paul Doignon and Jean-Claude Falmagne made similar contributions on mathematical models of knowledge spaces. Here the "knowledge state" of an individual represents the set of questions in a given domain, that the individual is capable to answer. In a quasi ordered set the "knowledge state" of an individual is somewhere between complete ignorance and total knowledge. This approach can be used to model knowledge assessment, communication processes and learning processes.

The applications level of the model is very important, because here we describe the fundamental aspects of communication and the way partners may use the documents within their applications (Reusch: Universale Kommunikationsmodelle ..).

4. XML-BASED IMPLEMENTATION OF A MULTI-LEVEL COMMUNICATION MODEL

The implementation of such a model today should be based on XML. Here it is easy to integrate all kinds of data and metadata and to link data.

Several groups work on "Meta Content Framework Using XML" (http://www.textuality.com/mcf/NOTE-MCF-XML.html).

The application of XML-based documents grow rapidly and even XML-based thesauri for the content description of documents are still available. One XML-based thesaurus was derived within the project HyperLib of the Loughborough University (UK) and the University of Antwerp (B) (http://lib.ua.ac.be/docstore.html). HyperLib metadata are partly defined by a Document Type Definiton (DTD) according to XML-standards. The following figure shows the main elements of the definition of the head of such a thesaurus.

More important is the description of the body, where keywords in several languages and links to related keywords (synonyms, related terms, …) must be handled.

The following figure shows the main elements of the definition of the body of such a thesaurus.

```
<!ELEMENT   thesaurus (head?, body)>

<!ELEMENT    head (title, manager, author+,copyright, notice)>
<!ELEMENT    title (#PCDATA)>
<!ELEMENT    manager (#PCDATA)>
<!ELEMENT    author (#PCDATA)>
<!ATTLIST    author
          function (manager|editor|contributor) contributor>
<!ELEMENT copyright   EMPTY>
<!ATTLIST copyright
          year          number          #required
          company    CDATA          #required
<!ELEMENT notice    (#PCDATA)>
```

Fig.5 – *The HylerLib-Thesaurus-DTD – Head.*

```
01<!ELEMENT   thesaurus (head, body)>
02<!ELEMENT    body (introduction?, keyword+)>
03<!ELEMENT    introduction (#PCDATA)>
04<!ELEMENT    keyword (v+)>
05<!ATTLIST    keyword
06          name            CDATA          #REQUIRED
07          source          CDATA          #IMPLIED>
08<!ELEMENT    v (n)>
09<!ATTLIST    v
10          type  (main|syn|rt|bt|nt) syn
11           language ( ger|eng|fr|nl|rus|pl) ger>
12<!ELEMENT    n    (#PCDATA, s*)>
```

Fig. 6 – *The HylerLib-Thesaurus-DTD – Body.*

The body may include an introduction (line 02) followed by any number of keywords. The name of the keyword is defined as attribute of the element keyword. Language specifications and links to related keywords are implemented using any number of elements v each of which has as single sub-element n.

The following figure shows a concrete keyword based on this DTD.

```
<thesaurus> ..
<head>  ....  </head>
<body> ....
<kw name="journal">
<v type="main" language="eng">          <n>journal
<v type="main" language="deu">          <n>Zeitschrift
<v type="main" language="nl">           <n>tijdschrift
<v type="syn"   language="nl">          <n>periodiek
<v type="syn"   language="eng">         <n>periodical
<v type="syn" language="fr">            <n>pйriodique
<v type="nt"   language="eng">          <n>yearbook
<v type="nt"   language="deu">          <n>Jahrbuch
<v type="nt"   language="nl">           <n>jaarboek
<v type="nt"   language="eng">          <n>annual report
```

Fig. 7 – *Keyword according to the HyperLib-DTD.*

The keyword has a unique name. The equivalent names of that keyword in the languages used are defined by sub-elements v of type main with the attribute language and an attribute value describing the concrete language. The "name" of the keywords in the v-element is defined as attribute of the n-sub-element of the v-element.

The keywords that are not "main"-keywords are v-elements of type syn (synonym), nt (narrower term) and so on.

This approach is open for several kinds of metadata that can be used separately or combined. The following figure shows an approach with keyword classes. One class may be used to desribe resolutions, another to desribe law, another to describe documents dealing with tourism.

```
01<!ELEMENT   thesaurus (head, body)>
02<!ELEMENT    body (introduction?, keywordclass+)>
03<!ELEMENT    introduction (#PCDATA)>
04<!ELEMENT    keywordclass (keyword+)>
05<!ATTLIST    keywordclass
06          Classname       CDATA          #REQUIRED
07          Classtype (resolution|law|tourism) resolution>
08<!ELEMENT    keyword (v+)>
09<!ATTLIST    keyword
10          name            CDATA          #REQUIRED
11          source          CDATA          #IMPLIED>
12<!ELEMENT    v (n, excl*,incl*)>
13<!ATTLIST    v
14          type  (main|syn|rt|bt|nt) syn
15           language ( ger|eng|fr|nl|rus|pl) ger>
16<!ELEMENT    n    (#PCDATA, s*)>
17<!ELEMENT    s    EMPTY>
```

Fig. 8 – *Thesaurus-DTD with keyword classes.*

This is the first step to a multi-level communication system based on XML supporting several kinds of metadata. The transformation of the DTD to an XML-schema should follow with more advanced links. XSL-modules can used to define information retrieval processes.

6. REFERENCES

[1]Baeza-Yates, Ricardo; Ribeiro-Neto, Berthier: Modern Information Retrieval, Reading 1999.

[2]Ganter, Bernhard; Wille, Rudolf: Formale Begriffsanalyse, Berlin 1996.

[3] Lindig, Christian: Algorithmen zur Begriffsanalyse und ihre Anwendung bei Softwarebibliotheken, Dissertation Universitдt Braunschweig 1999.

[4]Reusch, Peter J. A.: Universelle Kommunikations-modelle zur Unterstьtzung des E-Business, Conference of the Academy, Samara 2001.

[5]Reusch, Peter J. A.: Universelle Kommunikations-modelle zur Unterstьtzung des E-Business, Dortmund 2001.

[6]Reusch, Peter J. A.: Wissensmanagement, Textbook, planed for autumn 2001.

[7]Reusch, Peter J. A.: Data Driven Data Reduction for Data Analysis and Knowledge Discovery, International Workshop on "Discrete optimization methods in scheduling and computer-aided design", Minsk 2000.

[8]Reusch, Peter J. A.: Modellverwaltung und Expertensystemkomponenten für betriebliche Infor-mationssysteme, Mannheim 1988.

[9]Reusch, Peter J. A.: Informationssysteme, Dokumentationssprachen, Data Dictionaries, Mannheim 1980.

[10]Wille, Rudolf; Zickwolff, Monika: Begriffliche Wissensverarbeitung, Mannheim 1994.

[11]Stumme, Gerd; Wille, Rudolf: Begriffliche Wissensverarbeitung, Methoden und Anwendungen, Berlin 1999.

[12]W3C: Resource Description Framework (RDF), www.w3c.org/TR/rdf-schema/

[13]Doignon, Jean-Paul; Falmagne, Jean Claude: Knowledge Spaces, Berlin 1999.

Prof. Dr. Dr. h.c. Dr. h.c. Peter J. A. Reusch

Present Main Position

Professor for Management and Information Sciences
Fachhochschule Dortmund – University of Applied Sciences
Dortmund – Germany

Secondary Positions

Belarussian State Economical University - Minsk
University of Latvia - Riga

Curriculum Vitae (1.6.2002)

1950 born in Bruehl near Cologne, Germany, on 26th of September

1960 – 1969 Apostel-Gymnasium Cologne, with specialization on Humanism

1969 - 1974 ***Studies at the University of Bonn****: Courses in Mathematics, Computer Sciences, Management, Economics*

1974 ***Diploma in Mathematics*** *Thesis: Complexity of Elementary Functions*

1974 - 1976 ***Post-graduate Studies at the Universities of Bonn and Utrecht*** *Cources: Information Retrieval, Algebra, Developmental Systems, Fuzziness, Management*

1976 ***Doctoral Disseration at the University of Bonn*** *Thesis: Generalized Lattices as Fundamentals to Retrieval Models, Multidimensional Developmental Systems and the Evaluation of Fuzziness*

1977 - 1988 ***Head of the Department for Information Systems*** *at Rheinische Braunkohlen Werke AG in Cologne – German Brown Coal Mining*

1981 - 1991 ***Lecturer at the University of Bonn*** *– Information Systems – part time*

1988 - now ***Professor for Management and Computer Science*** *at the Fachhochschule Dortmund – University of Applied Sciences*

1991 - now ***President of the Gesellschaft*** *für betriebliche Informationssysteme und Expertensysteme mbH - IBIES*

1991 - 19962001 - now ***Dean*** *of the Faculty for Management and Economics at the Fachhochschule Dortmund – University of Applied Sciences*

1991 - 1997 ***Professor at the University of Stettin****, Poland – part time*

1995 ***Doktor oek. nauk h.c.*** *of the Belarus State Economic University, Minsk*

1996 ***Doktor oek. nauk h.c****. of the University of Latvia, Riga*

1999 - now ***Professor at the Belarus State Economic University in Minsk*** *– part time*