Konrad-Zuse-Zentrum
für Informationstechnik Berlin

Takustraße 7
D-14195 Berlin-Dahlem
Germany

MARTIN GRÖTSCHEL, JOACHIM LÜGGER

Scientific Information Systems and Metadata

Preprint SC 98-26 (October 1998)

Scientific Information Systems and Metadata

Martin Grötschel **Joachim Lügger**

Abstract: This article begins with a short survey on the history of the classification of knowledge. It briefly discusses the traditional means of keeping track of scientific progress, i.e., collecting, classifying, abstracting, and reviewing all publications in a field. The focus of the article, however, is on modern electronic information and communication systems that try to provide high-quality information by automatic document retrieval or by using metadata, a new tool to guide search engines. We report, in particular, on efforts of this type made jointly by a number of German scientific societies. A full version of this paper including all hypertext references, links to online papers and references to the literature can be found under the URL: `http://elib.zib.de/math.org.softinf.pub`

Introduction

The exponential growth of information, in particular in the sciences, is a topic discussed broadly. The problems arising by this increase are treated in depth in Odlyzko (1995). This development cries for adequate organization of knowledge and for efficient means of retrieval of information. The information flood is fostered by the tools provided by the Internet. At the same time, these technological developments seem to make efficient information handling feasible. This will be discussed in this paper.

Before we do that let us see how previous generations have coped with the problems of making knowledge accessible. The most prominent approach was "classification". The uninitiated observer may believe that classification was nothing but a way to organize knowledge so that relevant information can be found easily. The claim *"Classification is power"* may thus come as a surprise. However, this is nothing but our concise synopsis of the article Darnton (1998). It also follows from a combination of John Dewey's claim *"Knowledge is classification"* and Francis Bacon's *"Knowledge is power"*.

Let us explain these statements. Whenever organized information is offered an explicit or implicit classification is used. In the Internet organized information is provided through, e.g., a universal virtual library such as Yahoo!, a subject-specific list of links like Math-Net-Links, a search engine such as GERHARD (see Koch et al. (1997) for URLs and more detailed information about the role of classification schemes in the Internet). Why is such a classification the execution of power? The key observation is that the classifier decides which topic is important (high on the list, or on the list at all); the search engine designer does the same through the rules of his ranking algorithm. He may manipulate the world by listing information first he likes best (or for which he gets paid) in the same way as an encyclopedist focuses the attention of a reader along the lines and branches of his design of the "tree of knowledge". Darnton (1998) outlines this aspect with respect to Diderot's *Encyclopédie.*

D'Alembert (1751/1998) discusses Diderot's plans to design a world map of knowledge that can be navigated easily. He also justifies their joint decision to abandon this plan because one could *"think of as many scientific systems as world maps of different views, whereby each of these systems has a specific exclusive advantage in favor of the others"*. They observed that all arrangements of knowledge are arbitrary, and that each has a great number of inherent defects and unsolvable contradictions, in particular, if the evolution of knowledge over time is taken into account. Thus, they decided for alphabetic ordering, nowadays called *lexicographic index*, which is the traditional way of "implementing" information retrieval. In fact, 250 years later the developers of Alta Vista, one of the most successful Internet search engines, were confronted with exactly the same problem, namely "to index or to classify", see Seitzer et al. (1997). They decided for indexing and information retrieval and against a hierarchic classification, e.g., as employed by Yahoo!.

The strengths and weaknesses of classification and information retrieval are widely discussed, in particular by those who dream of a "universal, heterogeneous, world-wide digital library". It follows from d'Alembert's observation that a single universal classification scheme or an information retrieval mechanism alone do not suffice to create such a library. The new idea is to furnish data with additonal data about these data, called *metadata*, that allow to view the world from different perspectives.

Although the initial goal of the "metadata move" was to help the authors of web resources to make the results of their work more visible, the current development aims at more general goals. Both, the attributes and the contents of metadata are still in the design process. The initiative is fostered not only by the providers of digital libraries. It gained momentum by a broad acceptance within the group of the more traditional contents providers (publishers, museums, libraries, archives, document delivery services, etc.). Thus, it now also aims at providing suitable metadata for all kinds of traditional documents and for the more complex digital items in the web (which do not only include books and papers, but also videos, music, multimedia information, hypertexts), now termed *document-like objects.*

Before stepping into the details of metadata let us briefly review some of the developments, relevant for the topic, that took place before the birth of the web.

1 Classification in the Sciences

Metadata have been invented to form a basis for navigation in data sets of large scale. They have not been conceived in the context of classification. Classification systems have been designed for structuring the body of knowledge so that new information can be incorporated easily and already archived information can be found effectively. Although metadata and classification systems aim at completely different targets, there is an intimate relationship. We want to explore this briefly. We start with a short sketch of the history of classification in the sciences.

Classification is mainly motivated by two objectives:

- to introduce structure into masses of facts,
- to build a "unified and homogeneous" view of the "world".

History bore witness for some of the potentials of classification.

Lenné's classification of plants and animals started biology as a science in the 18th century, see Rossi (1997). Lenné's system is still in use today. In fact, Lenné can be viewed as the father of modern taxonomy. To name another example, modern chemistry started as a science with the periodic system, which was "found" in the middle of the 19th century after a number of false starts, see Bensande-Vincent (1989).

Designers of classification systems always tried to follow three principles. Their system should be organic, simple to grasp, and simple to memorize. By this design, such systems can be viewed as communication tools. In fact, the "librarian" Melvin Dewey used numbers to name the classes in his system and, thus, imaginated his Decimal Code (DDC) as a language-independent universal communication system, see, e.g., Dahlberg (1974).

The early designers of classification systems in the 18th century also conceived knowledge as a new territory to be discovered. They wanted to provide tools for navigation in this unknown landscape. (Notice the similarity to "navigation concepts" in the modern World Wide Web.)

Darnton (1998) states that *Mappemonde* was a metaphor central to Diderot and many other encyclopedists. Diderot viewed his *Encyclopédie* as a world map showing the connections and interdependencies among the most important *countries*.

An elementary feature of a classification system is that it draws borders. It must do this in order to distinguish objects. But borders, wherever they are, are dangerous. They are subject to attacks. Here attacks come from different views or from new knowledge. If too many of the important borders go, a system disintegrates. This danger causes fear and, in turn, rigidity. Lynn Margulis, for instance, describes this phenomenon. She created a new theory of the origins and evolution of cells, summarized in her book *Symbiosis in Cell Evolution*. Her analysis strongly impacts on biological taxonomy and systematics. She describes in her paper (Margulis (1995)) the rigidity of the establishment that was very reluctant to accept the new view. Among many other obstacles she lists: *"... a school or publisher would have to change its catalog. A supplier has to relabel all its drawers and cabinets. Departments must reorganize their budget items, and NASA, ..., and various museums have to change staff titles and program-planning committees. The change ... has such a profound implication ... that resistance to accept it abounds .... It is far easier to stay with obsolete intellectual categories."*

Margulis' remarks apply in general. Scientific progress is a danger for every classification system.

Thus, whatever genius is used for their design, classification systems are limited in range and in time, often inconsistent and illogical and, partly, even contradictory or paradox. This holds, in particular, if ad-hoc adaptations are made to incorporate new developments. This is one of the reasons for Daston's statement: *"Classifications organize, but they are not organic."*, Daston (1997).

For instance, biology before Darwin was organized according to the "rule of five", which was supported by many leading scientists of this time, see Gould

(1985). It is almost unimaginable that a reader of our time could "believe" in such a system. It took a revolution to change this view.

Why have we told these stories about traditional classification systems and their disadvantages? Well, because one can observe that history seems to repeat itself in the world of electronic information. With the advent of powerful computers, cheap storage devices, and fast networks (in short: with the rise of the Internet) a flood of electronic information appeared. Catchwords such as "information society" were quickly coined pointing at the fact that in the electronic world digital information is accessible from everywhere, by everybody, at any time. However, it was soon realized that Internet information is "chaotic". To cure this desease classification systems came up quickly. According to Koch (1997) there are a number of classification mechanisms operating in the World Wide Web. They provide for

- Support for navigation by
  - structuring for browsing (e.g., *WWW Virtual Library*),
  - setting context for searches (*Scorpion*),
  - broadening/narrowing searches (*Yahoo!*),
  - help to master large databases (*MSC index*).
- They also offer support for communication, by
  - organizing large sets of electronic discussions (UseNet News),
  - presenting/accessing knowledge in a common (uniform) way,
  - allow for interoperability of databases ("crosswalks"),
  - stabilize contexts and conceptual schemes for distributed user communities in networks.

In the electronic world classification systems suffer from the same deficiencies as they do in the traditional world.

In fact, some of these deficiencies become even more visible. It is generally agreed that, fostered by the electronic revolution, progress in the sciences and the production of information is becoming faster and faster. Rapid changes make the rigidity, fixed granularity, and slow adaptivity of classification systems very apparent.

On the other hand the electronic world offers new opportunities that significantly enlarge the power of traditional navigation provided by classification systems. For instance, modern hyptertext systems in the World Wide Web offer graphical and spatial navigation by means of

- Interactive maps without any limit on the depth of nesting (e.g., *Virtual Tourist*, *CityNet*)
- Combination of pictures from the earth (or even space) with geospatial coordinates (*EarthView*, *Living Earth*)
- a list of icons to select from collections of video clips (*Cine Base Video Server*)
- two-dimensionally arranged collections of icons representing maps producing, when selected, regularly updated geospatial information such as weather forecasts, temperature maps, etc. (*Blue-Skies Weather Maps*)

An ever growing spectrum of alternative navigational paradigms is getting into common use today, such as

- Navigation by historic terms, chronologies or history maps (*History of Mathematics* from *Mac Tutor*, *Chronology of Mathematicians*)
- Navigation by theory, e.g., by mathematical expressions (*Famous Curves Index* from *Mac Tutor*)

We expect that the full spectrum of document/resource description and related navigational facilities – as they are in use already in modern hypertext systems, like Hyperwave, see Maurer (1996) – will come to the Web with the new extended markup language XML, which is based on SGML and supported by the World Wide Web Consortium W3C, see Mace et al. (1998).

2 To Index or to Classify

One of the severe drawbacks of classification systems is that they have to be supported by manpower. A group of persons must agree on the interpretation of the items of the scheme. Information must then be processed by a person who evaluates the contents, selects key words, and assigns the objects to appropriate classes of the scheme. Thus, such systems have limits due to the availability of time, manpower, and financial means. In general, classification systems cannot keep up with the growth of knowledge; in modern terms, they don't scale.

In fact, d'Alembert did not only observe this phenomenon, he also argued, as mentioned before, that the adaptation of a classification system is outpaced by the growth of information. Therefore, d'Alembert and Diderot decided against classfication and arranged their *Encyclopédie* lexicographically, i.e., they decided to index and not to classify.

Those groups of people aiming at making web information more accessible were confronted with the same problems as d'Alembert, however, at a much larger scale.

The first efforts to organize web information were based on alphabetic lists of web resources; the most prominent example is the WWW virtual library. Although this is a very valuable resource locator and although its maintenance is distributed on many shoulders it cannot keep up with the growth of the web. In fact, similar endeavours, such as the Geneva's University GUI catalogue, stopped their service. Considering these difficulties it is apparent that one has to look for "automatic solutions". Considerations of this kind gave birth to "search engines". One prominent example, among many others, is Alta Vista. The goal of the designers of Alta Vista was to index the contents of the whole (accessible) web. The only way to establish and maintain such an index at acceptable costs is to use so-called "robots", i.e., programs that traverse all available web resources (by following all the links they can find), extract, index, and rank all "relevant" information they can find and that concentrate the results into one huge data base served by very powerful computers. In the beginning of these projects it was quite unclear whether search engines would be able to achieve their ambitious goals. Today they have become a tremendous success. Basically everybody using the Internet employs search engines to find information.

A different path was taken by the designers of the (now) commercially very successful Yahoo!. Yahoo! has developed its own classification system (they call it *ontology*) and employs a group of classifiers (currently about 20) who select information resources that they view valuable for the "Yahoo! customers". These resources are classified (the staff writes short descriptions) and integrated into the Yahoo! scheme. Whatever is contained in the Yahoo! system can be retrieved using a specific search engine. Thus, Yahoo! is a combination of a standard (but new) classification system that is based on handcraft (evaluating/ranking by a group of experienced classifiers) with modern electronic retrieval tools. The limitations on manpower force concentration on special topics and strict selection. What may seem weakness has turned into strength since the customer of Yahoo! is sure to obtain information assessed by competent persons.

The obvious question is: Can't one replace the experienced classifiers by automatic evaluation systems? This question has been asked more than thirty years ago and gave rise to the theory of information retrieveal. Among the key terms in this theory are "precision" and "recall". (*Precision* is the number of retrieved and relevant items divided by the number of all retrieved items. *Recall* is the number of retrieved and relevant items divided by the number of all relevant items.) It seems that these terms provide good tools to describe the quality of the answers an information retrieval system gives. However, both definitions contain the term "relevant", which is not a technical term but a "concept of mind". Confronted with a larger collection of, e.g., scientific papers even a specialist does not know which papers are "relevant" for him. How should a machine do so? Thus, there are small chances to get help from computers in analyzing "relevance" in document collections of significant size. Furthermore, what is relevant may change over time or after having obtained new information.

Blair (1990) discusses in depth why relevance is difficult to ascertain, and that measuring the success of retrieval results is difficult and very costly. He argues that this is almost impossible for large databases. Research in information retrieval stalled about fifteen years ago. However, interest in information retrieval techniques is rising again. This rise is not only fostered by the appearance of search engines and the growth of the Internet, but also by the world of "large-scale research". For instance, in the Human Genome Project, massive sets of data are produced (by making automated experiments and automatically measuring the results) that must be recorded, linked to and combined with other data. These data must be classified along the prevailing theories and connected to associated publications. Furthermore, statistics, visualizations, etc. have to be produced.

While traditional classification and information retrieval systems are based on a linguistic approach (organizing knowledge, uniform view, universal requirements, and document retrieval) the new demands focus on pragmatic topics (organizing documents, user-specific needs, adaptable views, and data retrieval). In fact, this very same change of view was the incitement for the digital library projects in the United States and the United Kingdom.

The concept of metadata that we will discuss in the next section arose within these digital library projects.

3 Towards Resource Discovery in Networks

With the advent of the *Datenautobahn*, the Internet and its large and widespread digital resources we are confronted with yet another order of complexity. The big Internet archives not only contain text material (like preprints and electronic books), they also include images, maps and geospatial data, videos and computer vision material, environmental and agricultural databases, vast arrays of governmental and statistical data, pictures from the universe, etc.

Right here, on the information highway, library science, communication, and computing are merging. Thus, it is no wonder that the origins of metadata are rooted in the digital library projects supported by NSF, NASA, and ARPA. Nevertheless, a well known document deliverer, OCLC, and a Supercomputing Center, NCSA, have started the first concrete "universal" metadata activity. They perceived the Dublin Core, see Weibel (1995) for a report on the first workshop in Dublin, Ohio, where the term *Dublin Core* was coined. Later, UKOLN, the UK Office for Library and Information Networking of Great Britain, and many other working groups, user communities and organizations joined the project, e.g., national libraries, museums and institutions of cultural heritage.

A new (very pragmatic) paradigm came up with this movement: usability and utilization, instead of knowledge ordering and information retrieval. To say it short: the focus is on data – not on knowledge and not on information. And, there are "data about data" from now on called *metadata*, conceived as "information that makes data useful". This concept is centered around the user and his needs. There is another shift. The user is not only a consumer who wants to discover resources in the Internet, the user also offers his resources and is asked to do so.

3.1 Dublin Core, Issues and Problems

The Dublin Core is still in active development. In the beginning (spring 1995), as Weibel et al. formulated in the OCLC/NCSA Metadata Workshop Report, *"The discussion was ... restricted to the metadata elements for the discovery of what we called document like objects, or DLO's by the workshop participants"*. The Internet was considered chaotic and the proposed solution was to provide authors of Internet resources with metadata techniques from the library and information sciences. They should, however, be easier to use. The initial aims of the designers were very ambitious. The Dublin Core elements should guarantee:

- Intrinsicality,
- Extensibility,
- Syntax independence,
- Optionality,
- Repeatability,
- Modifiability.

At that time (and still today) a great number of different description schemes for resources were in use in different user communities, see Dempsey and Heery et al. (1997) and Heery (1996) for an overview of present metadata formats. The question was, do these have something in common that would help the authors of Web resources? The answer given at the first Dublin Core Workshop (we will use the abbreviations DC and DC-1) was a core set of twelve elements, the elements numbered 1, ..., 12 in Table 1. DC-3 added three more elements. All these were grouped and finally named (the current usage is indicated in bold) at DC-5 in the way shown in Table 1 (`http://purl.org/metadata/dublin_core_elements`).

	Content		Intellectual Property		Instantiation
1.	**Title**	2.	Author or	7.	**Date**
3.	**Subject** and		**Creator**	8.	Resource
	Keywords	5.	**Publisher**		**Type**
4.	**Description**	6.	Other	9.	**Format**
11.	**Source**		**Contributor**	10.	Resource
12.	**Language**	15.	**Rights**		**Identifier**
13.	**Relation**		Management		
14.	**Coverage**				

Table 1: The Dublin Core Metadata Element Set, according to DC-5

In the meantime five major Dublin Core workshops took place, the last one, DC-5, in October 1997 in Helsinki. There was a shift in major principles as well as in the constitution of the DC community, both due to wide international discussion and broad acceptance of the general idea. According to the DC-5 Report given by Weibel and Hakala (1998), the design principles are now

- Simplicity,
- Semantic Interoperabiltiy,
- International Consensus,
- Flexibility.

The center of gravity in the activities of the DC community has changed from authors (laymen) to catalogers (information professionals). As Priscilla Caplan (1997) reports in the PACS-review: *"Back in 1995 we focused on providing authors with the ability to supply metadata as they mounted their own publications to the Web. This is happening, but not as much as we expected; most metadata is being created by catalogers, or information professionals we wouldn't quite call catalogers, or by other non-authorial agents."*

Today, technical rather than conceptual questions have moved into the focus of the dicussions, e.g., integration of heterogeneous databases, interoperability of digital libraries, combinations of digital resources, and wide accessibility of catalog information. The DC community encompasses a broad spectrum of groups from libraries, museums, archives, documentation centers, both public and commercial, and also a number of scientific groups, e.g., from mathematics, astronomy, geology, and ecology. These groups have agreed on extending and applying the DC metadata principles to non-textual objects (e.g., scanned images, digitized music, and videoclips) and also to non-Web objects such as entries of library OPACs, catalog information on visual arts and historic artefacts, which cannot be scanned at all.

In spite of these substantial extensions in aims and targets the Dublin Core remained simple. It is a conceptual scheme which – free from the peculiarities of syntax and implementation – can be described by no more than three typewritten pages. The DC community strongly supports implementation projects in the World Wide Web (based on HTML) and in the Z39.50-oriented database community. The Helsinki Workshop web pages list about 30 major implementation projects (`http://linnea.helsinki.fi/meta/`), e.g., the Math-Net project of mathematics in Germany.

The DC community gained momentum through its ability to integrate a variety of scientists and cataloging people, who are creating their own metadata methodologies according to their special habits, needs and uses, and who are increasingly realizing that information exchange between the sciences and society is becoming more and more essential. For them, the Dublin Core provides a "window to the world".

This picture describes the situation quite precisely. One can view the world of information as a set of many rooms containing massive sets of heterogeneous data, in general not accessible by inhabitants of other rooms. DC metadata provide a uniform interface through which search engines, robots, etc. can collect information about the data in other rooms. The search engines etc. obtain the ability to gather and process the attributes and allow the viewer to inspect them in an integrated environment. This supports inter- and transdisciplinary information exchange far beyond what traditional libraries can offer. This is in line with the growing trend in the sciences to present results to the general public.

3.2 What the Dublin Core cannot do, and ways out

Bibliographic cataloging, in a few words, consists of a set of rules by which information about a book can be reduced to a catalogue card in a systematic way. To make this work, throughout the world, rule systems, such as RAK in Germany or AACR in the United States, have been designed that provide very good guidelines for the professional cataloger but are far too complex for the "educated layman". One of the ideas behind the Dublin Core was to extract the "best of this world" so that authors of web objects can describe their products without professional aid. The products for which the Dublin Core has been designed are what was called "document-like objects", which may be everything that is stored electronically in the web, e.g., electronic versions of books and journals, digital maps, sources of programs, geospatial or medical data, etc.

Of course, the communities working with computer programs, medical data, etc. have also developed their own description schemes, and they use differently constructed data bases. It was, thus, another aim to formulate the DC concept in such a way that also the central attributes of these descriptions can be included. Interoperability of the respective technical system was a main goal.

If you are determined to stay "simple and universal", as the Dublin Core does, you cannot describe everything. Moreover, in the implementation process there is no way to escape from specifying details. This was also apparent to the DC designers. At the second workshop, DC-2 in Warwick, UK, a conceptual framework, called the *Warwick container architecture*, was created, which allows to support and enrich the Dublin Core by sets of additional description elements, see Weibel and Hakala (1998) for a review of DC-1 to DC-5. A detailed development of the Warwick framework has, however, not happened yet. The group working on this issue has joined its forces with another group working on a similar topic in developing the Resource Description Framework (RDF) for the World Wide Web. This is based on XML (eXtended Markup Language) that is viewed, by almost everyone involved, as one of the future web languages. The DC community also made other steps to integrate large potential user groups.

A surprising result of DC-3, the CNI/OCLC Image metadata workshop, which took place in Dublin, Ohio, in September 1996, was that now non-Web objects can also be treated adequately within the Dublin Core. As a consequence, the DC community receives useful support and criticism also from the (traditional) cataloging community. This would not have happened without the inclusion of the 15th element, Rights Management, because visual art and digitized images are often affected by copyright regulations, as are data bases with specialized information.

The DC-3 workshop also made the limitations due to the restriction on only 15 attributes of the DC concepts clearly visible. This led to some tension within the DC community, which were partially resolved at DC-4 in Canberra, Australia, in March 1997. It was accepted as a solution that DC metadata should be enhanced by (at least) three qualifiers in order to get more expressive power. The so called 3 *Canberra Qualifiers* are: LANG (to characterize the language a specific metadata element is written in), TYPE in the meantime called SUBELEMENT (to specify subfields for greater precision), and SCHEME (to specify a bibliographic scheme or international standard used). Each of these qualifiers is under development in different DC working groups.

To give some examples, we will now discuss a few problems (out of a broad spectrum) that became visible through the experience of a number of implementation projects; for details see the extensive discussions in the DC meta2 mailing list (meta2@mmrl.ut.ac.uk).

You will need the LANG qualifier in order to write the title element

DC.Title = (LANG = ...) ...text ...

of a resource in any case you are not using the English language (which is the default) for the text.

And you will need a subelement, e.g.

DC.Title.Alternative = ...text ...

for any title other than the main title, where "title" is the name of the resource, usually given by the creator or publisher.

The Creator (or Author) element of a resource is packed with problems once you start to think about it. You need (the help of) a scheme to write (and search for) it correctly. How would you code the name of the author? Which one of the following alternatives would be correct?

DC.Creator = Grötschel, Prof. Dr. M.
DC.Creator = Prof. Dr. Martin Grotschel
DC.Creator = Martin Groetschel
DC.Creator = Grötschel, M., Prof.

You must write a name "correctly", if you want to have reasonable alphabetic lists, for instance. You will also need a coding convention for accents, umlauts, etc., e.g., to sort names consistently. Professional catalogers and librarians are using name authority files, such as the LCNAF (Library of Congress Name Authority File) from the LOC; in Germany one would use PND, the PersonenNamenDatei, or GKD, the Gemeinsame Körperschafts-Datei.

DC.Creator = (SCHEME=LCNAF) Grötschel, Prof. Dr. Martin.

Apart from LCNAF, PND, etc. there are many other kinds of "controlled vocabularies". If you think of subjects and keywords or classification codes; there are just a few: LCSH, MeSH, AAT, LCC, DDC, UDC, BC, NLM, MSC.

You will need subelements also for proper discrimination in searches, e.g., in specifying for greater precision in search:

DC.Creator	=	...
DC.Creator.PersonalName	=	...
DC.Creator.CorporateName	=	...
DC.Creator.PersonalName.Address	=	...
DC.Creator.CorporateName.Address	=	...

But who is the creator of a digitized painting by Picasso? Is it the person who digitized it (and put it in the Web) or is it Picasso? An answer to this question was given at DC-5 in Helsinki by means of the

1:1 Principle: Each resource should have a distinct metadata description and each metadata description should include elements for a single resource. It is desirable to be able to link these descriptions in a coherent and consistent manner by usage of the RELATION element.

The consequences of this decision are not yet fully understood. The relation field is under development and will go through some major evolution in the near future. At present about five major types of relations are discussed in the relation working group:

1. Inclusion relation (e.g., collection, part of)
2. Version relation (edition, draft)
3. Mechanical relation (copy, mirror copy, format change)
4. Reference relation (citation)
5. Creative relation (translation, annotation)

If all these problems connected with names are solved, then there remains the (what we call) "Tschebytscheff Problem". As M. Hazewinkel (1998) pointed out on the occasion of a metadata workshop in Osnabrück, Germany, there are more than 600 variants of writing Tschebytscheff, the name of a famous mathematician, correctly.

We agree with Mary Lynette Larsgaard (1997), a spatial-data cataloger at the Map and Imagery Laboratory, Davidson Library, University of California, Santa Barbara: *"Full cataloging is a complex, time-consuming process. Library administrators, when they feel like being horrified, figure out how much time (and therefore money) it takes per title – around $ 67 per item, at least at Davidson Library, ..."* and *"There are many more possible methods of access where full cataloging is used; the question is, how necessary are they? And the answer is, it depends. What are users looking for?"*

But, summarizing her experience in cataloging images from the Web using Dublin Core elements, she also states: *"The general experience in university libraries is that a brief record is sufficient, and indeed, this brief record is what normally displays in a library online catalog. Only the place of publication does not appear in the Dublin Core element set."*

3.3 Metadata and Classification

Classification systems can be categorized, according to T. Koch (1997) into

- Universal schemes (e.g., LCC, DDC, UDC)

- National general schemes (e.g., BC/PICA, RVK)
- Subject-specific schemes (e.g., MSC, NLM)
- Home grown schemes (e.g., Yahoo!'s anthology)

Universal schemes are ponderous, partly contradictory, and they are not well known to the scientist. Their advantage is their potential for "normalisation", e.g., by providing a framework for controlled keywords. In fact, this is their main use in universal libraries. Subject-specific schemes, in contrast, are in frequent use within certain scientific communities. They often utilize them for communication purposes. Subject-specific schemes, however, rarely transcend the borders of the specific community. National general schemes, on the other hand, are limited by their inherent range of acceptance. Home grown schemes, finally, may be accessed and used worldwide, but, unfortunately, they appear in general as the result of the activity of few persons or enterprises. Such schemes often disappear as soon as their creator gives up or lacks in commercial success.

Suppose there would be a universally accepted classification scheme and there would be vast sets of resources, perfectly classified according to the scheme, then we would have reached the heaven of search and retrieval. By dynamically adjusting the granularity of our search we could easily find those documents that match our interest. The world has not reached this state. Neither do we have a generally accepted classification scheme nor is all relevant information classified. This, in particular, holds for web documents.

We view metadata according to the Dublin Core scheme as a reasonable description of (web and non-web) documents and document-like objects. The Dublin Core elements constitute a conceptual description scheme that seems to form a good compromise between generality, precision, and simplicity. What is not so obvious is that it can also be used as a substitute for a universal classification scheme. In fact, special user groups can employ the Dublin Core to design and generate their own specific description systems.

To achieve this goal we have to assume the existence of a search engine that "understands Dublin Core", i.e., is able to restrict its search to the Dublin Core elements and allows to target searches onto words that are used as significant terms. This would result in a considerable improvement in precision without resorting to (enormously resource and time consuming) full text search. It would be desirable for all search engines to allow this option. At present, only a few experimental search engines of this type are existent.

If both, sufficiently many documents, described according to the Dublin Core scheme, and search engines understanding Dublin Core existed, the Web could be viewed as a "well-organized" global digital library. A key point here is that "well-organized" is not defined universally by some enlighted general committee; user groups (large or small) with certain common interests have to get together and to agree on their own standards, the usage of words, term hierarchies etc. to define what (within their local framework) well-organized is intended to mean. This results in a decentralized system where contradictions and conflicts may occur but that also has the potential to lead to globally accepted standards. We describe attempts of this type in the next section.

3.4 Efforts of Scientific Societies

In the early nineties the Bundesministerium für Forschung und Technologie (BMFT, now BMBF) supported projects by the Deutsche Mathematiker-Vereinigung (DMV) and Deutsche Physikalische Gesellschaft (DPG) to intensify the use of the mathematics and physics data bases at Fachinformationszentrum Karlsruhe within the academic community. The participants of these projects soon realized that the use of some data bases is important but that the evolving Internet offers enormous potentials for electronic information, communication, publishing, etc. to support research and teaching. Moreover, it was obvious that most of the organization and planning to be done by the mathematics and physics societies is not subject specific. Since there were many overlaps and joint interests, it was decided to start a cooperative effort, called the "Gemeinsame Initiative der Fachgesellschaften zur elektronischen Information und Kommunikation" (short: IuK Initiative), and to join forces. Starting with the leading scientific societies in mathematics, physics, chemistry, and computer science, a treaty was signed, committees were founded, etc. to push the use of the Internet forward, to improve the computing and network facilities within the universities, to develop information systems, and so on, to make the electronic resources of the Internet more accessible to scientists and students "at their workplace". Even more important, all participating institutions and individuals were encouraged to make their own electronic resources widely available and "in an organized way".

Within this cooperation there were both discipline oriented projects, such as MeDoc in computer science (supported by BMBF) or Math-Net in mathematics (supported by Deutsche Telekom and DFN), and joint projects such as the Dissertation Online (supported by DFG). The IuK Initiative was intrumental in setting up GLOBAL-INFO, a BMBF-funded support program for global electronic and multimedia information systems. In all cases, emphasis was laid on interoperability, joint interfaces, international standards, etc. in order to be able to gain from the work of others.

At the same time the IuK Initiative realized that it had to act internationally. E.g., similar activities have started in other countries or, subject specific, on an international level. It is important for the IuK Initiative to coordinate its efforts with these activities to guarantee interoperability and provide mutual access to the respective information systems.

The IuK Initiative was successful in spawning many activities and projects, foster the development in other countries and internationally. Further leading societies in Germany from biology, education, psychology, sociology, electrical engineering, joined the initiative. The IuK Initiative cooperates with librarians, publishers, and other information providers.

Everybody involved in the IuK Initiative came to agree that, what is sometimes called the "information chaos in the Internet", has to be overcome, at least with respect to high quality scientific information. An important prerequesite for this is that information offered in the Internet is "well-structured". This is, however, not sufficient if one has in mind to automatically collect information, e.g., by means of web robots. Considering the amount of information offered in the Internet, automatic resource discovery is a must. This requires that resources are described by metadata. These metadata have to be produced manually, preferably by the authors who offer their resources. The metadata must be produced in a way that is understandable by robots. This way the Dublin

Core came into play. The DC initiative has, just as the IuK Initiative, broad transdisciplinary goals and is supported by a wide spectrum of scientists, catalogers, librarians, etc. This is why the IuK Initiative decided to play an active role in the general development of the Dublin Core and to start implementing the concept, e.g., with the Math-Net project involving almost all mathematics departments and research institutes in Germany. (And this is also why the authors of this paper got interested in this topic.)

3.5 Uses of Metadata: Final Remarks

As mentioned before, the development of metadata is not at all finished, especially, the Dublin Core is still undergoing technical and conceptual revisions. It is too early to judge whether this concept will be a success for the World Wide Web community as a whole.

Let us repeat, the Dublin Core is a conceptual framework for metadata formats. Each group using Dublin Core must specify, for each Dublin Core element, how to fill and interpret it. E.g., the element DC.CREATOR can be specified in various ways. The metadata set for a preprint in the Math-Net project uses DC.CREATOR for the authors of a paper. More precisely, they use the subfield DC.Creator.PersonalName, where the name of the author has to be written in the form "last name, first name ..." (without title). (The person inputing this data does not have to know technical details, such as coding a letter in HTML, since he only needs to fill out a simple form.) Other user groups employ DC.CREATOR (or subfields thereof) to specify the composer of a piece of music or a company owning a certain patent and will probably require different forms of notation. Catalogers who have been using bibliographic formats such as USMARC, MAB or the like are of course more familiar with such concepts. They typically define a "crosswalk" from their own data format to the elements of the Dublin Core by specifying, for each element, the related set of fields of their own format.

This indicates the complexity of the process. There will always be different user groups having their own interpretations of the Dublin Core elements. However, all interpretations are using the same Dublin Core format, the fifteen Dublin Core elements. This way the Dublin Core provides a bridge connecting the resources of many user groups. This observation supports Priscilla Caplan's view:

> *"Now it appears an even more common application of DC is as "lingua franca", a least common denominator for indexing across heterogeneous databases. ... the simplest way to index them all with some degree of semantic consistency may be to translate them all to DC."*

Dublin Core "is" in its range of elements a kind of "Inter-Meta-Data". To a certain degree it makes the integration of heterogeneous collections of resources possible. This is and will be more important in the future because of two reasons: (1) Inter- and transdisciplinary research projects are increasingly common in modern science. (2) The research process of today results in a variety of products, rather heterogeneous in form and contents (e.g., articles, books, software, large data sets, videos, etc.). If the Dublin Core will be widely accepted, also a market of search engines may evolve on the basis of future WWW protocol suits and employing the DC as universal data structure. Users of such engines may have access to an ever growing number of heterogeneous and well structured digital resources.

Our intention was to show where the Dublin Core and metadata described by Dublin Core elements can help organize knowledge that is reflected in data that are complex, heterogeneous, or interwoven.

Let us conclude – in analogy to Bearman (1995) – with a list of items, where metadata are absolutely necessary:

- Records with attributes of evidence:
  - Theses, dissertations,
  - Patents, authorship on new ideas,
  - Authentic art, orginality of work.
- Unique artefacts or protected items:
  - Collections of museums (bones, stones, ...),
  - Historical books/documents (papyri, ancient bible, ...),
  - Lecture notes, scientific books, audio cassettes, videos.
- Business environments, litigations:
  - Document management/delivery,
  - Online ordering.
- Archival, collections of statistics data:
  - Government, statistical authorities,
  - Local administrations.

In all of these categories the original data or artefacts cannot be "handed out" freely. In general, they must reside in an archive, a treasury, or a closed office, or in the depot, or a warehouse – until the moment where it is exposed, exchanged or sold. In all of these cases a certain substitute must exist which can be distributed freely or sent to a customer – instead of the original item. That is the purpose of all metadata.

References

BEARMAN, D. (1995): Towards a Reference Model for Business Acceptable Communications, in Richard Cox, e.d., *Recordkeeping Functional Requirements Project: Reports and Working Papers Progress Report Two*, University of Pittsburgh

BENSANDE-VINCENT, B. (1989): Mendelee: Die Geschichte einer Entdeckung. In: Elemente einer Geschichte der Wissenschaften, M. Serres (Hrsg.), Suhrkamp, 1994; Originalausgabe, Bordas, Paris

BLAIR, D. C. (1990): Language and Representation in Information Retrieval, Elsevier Science

CAPLAN, P.: "To Hel(sinki) and Back for the Dublin Core." The PACS Review 8, No. 4, 1997

DAHLBERG, I. (1974): Grundlagen universaler Wissensordnung. Probleme und Möglichkeiten eines universellen Klassifikationssystems des Wissens; Verlag Dokumentation, München

D'ALEMBERT, J. (1751/1998): Discourse préliminaire, (1751), reprinted e.g. by Fischer Verlag 1998; d'Alembert: Einleitung zur Enzyklopädie, Philosophische Bibliothek, 00473, ISBN 3-7873-1188-2

DARNTON, R. (1998): Philosophen stutzen den Baum der Erkenntnis: Die erkenntnistheoretische Strategie der *Encyclopédie*; in: Kunst & Geschichte, C. Konrad und M. Kessel (Hrsg.), Reclam, 209–241

DASTON, L. (1997): Die Akademien und die Einheit der Wissenschaften: Die Disziplinierung der Disziplinen; Manuscript of a lecture held at the Berlin-Brandenburgische Akademie der Wissenschaften, Dezember 1997, Max-Planck-Gesellschaft für Wissenschaftsgeschichte

DEMPSEY, L. and HEERY, R. (1997): Specification for resource description methods. Part 1. A review of metadata: a survey of current resource description formats. DESIRE project deliverable RE 1004 (RE), `http://www.ukoln.ac.uk/metadata/DESIRE/overview`

D-Lib Magazine, electronic version: `http://www.dlib.org/`

DEWEY, J. (1910): How we think, Boston, D. C. Heath

GOULD, S. J. (1985): Die Fünferregel. In: S. J. Gould, Das Lächeln des Flamingos, Betrachtungen zur Naturgeschichte, Suhrkamp 1995, 164-174; Originalausgabe bei W. W. Norton, 1985

HEERY, R. (1996): Review of Metadata Formats. Program: Autmated Library and Information Systems, Vol 30, Issue No. 4, October 1996

KOCH, T., DAY, M. et al. (1997): The role of classification schemes in Internet resource description and discovery: DESIRE (RE 1004) project deliverable, `http://www.ub.lu.se/desire/radar/reports/D3.2.3`

LARSGAARD, M. L. (1997): Metadata applied to Digitized Images in a Web Environment, Final Version: 7 Dec 1997; Map and Imagery Lab., Davidson Library, University of California,. Santa Barbara (`http://www.mathematik.uni-osnabrueck.de/projects/workshop97/papers/larsgard7.12.html`)

MACE, S., FLOHR, U., DOBSON, R. and GRAHAM, T. (1998): Weaving a Better Web; Byte (International Edition), Vol. 23, No. 3, March 1998, 58-68

MARGULIS, L. (1995): Gaia ist ein zähes Weibstück. In: Die dritte Kultur: Das Weltbild der modernen Naturwissenschaft, J. Brockmann 1998, 177-202, Originalausgabe bei Simon & Schuster 1995

MAURER, H. (1996): HYPERWAVE – The Next Generation Web Solution. Addison Wesley

NIELSEN, J. (1990): Hyptertext and Hypermedia, Academic Press

ODLYZKO, A. M. (1995): Tragic Loss or Good Riddance ? The impending demise of traditional scholarly journals. International Journal on Human-Computer Studies (formerly Intern. J. Man-Machine Studies) 42 (1995), 71–122

PACS-Review: Public Access Computer Systems Review, `http://info.lib.uh.edu/pacsrev.html`

ROSSI, P. (1997): Die Geburt der modernen Wissenschaft in Europa, Deutsch von M. S. Charmitzky und Ch. Büchel; Verlag C. H. Beck, München, Kapitel 14, 268–278; Originalausgabe bei Laterza, 1997

SEITZER, R., RAY, E. J., and RAY, D. S. (1997): The AltaVista Search Revolution, McGraw Hill

WEIBEL, S. (1995): Metadata: The Foundations for Resource Description, D-Lib Magazine, July 1995

WEIBEL, S., GODBY, J., MILLER, E. (1995): OCLC/NCSA Metadata Workshop Report, Office of Research, OCLC Online Computer Library Center Inc., and Advanced Computing Lab, Los Alamos National Laboratory

WEIBEL, S., HAKALA, J.: DC-5: The Helsinki Metadata Workshop. D-Lib Magazine, February 1998