Metadata in Science Publishing

Anita de Waard (Advanced Technology Group, Elsevier)
Molenwerf 1, 1014 AG Amsterdam

Joost Kircz (KRA-Publishing Research) *
Prins Hendrikkade 141, 1011 AS Amsterdam
kircz@kra.nl

Abstract

In the design of authoring systems in electronic publishing a great challenge is to what extent the author is able, can be enabled and is willing to structure the contribution her/himself. After all, all information that is denotated properly in the writing stage enhances the retrievability later on. Metadata are the crucial ingredients. Hence, prior to design and experiment is the need for a full-fledged understanding of metadata. In this contribution we discuss an attempt to classify metadata according to type and use and elaborate on the many complicated and unsolved issues. The message of all this is that metadata should be treated on equal footing as the objects they describe, in other words metadata are information objects in themselves. We show that all issues that pertain to information objects also pertain to metadata.

1. Introduction

With the impressive growth of hyper-linked information objects on the World Wide Web, the best possible way of finding gems in the desert is to create a system of filters - sieves, that enable a large throughput of information in the hope that the residue is of relevance to the working scientist. Two methodological directions can be taken to find relevant information. One approach starts from the assumption that information growth cannot be tamed. Purely statistical information retrieval techniques are a prime example of such an approach, which can be void from any semantic knowledge about the content at stake. In these IR techniques, context is inferred from patterns that contain the query words. In the extreme case, not even words are used as in the powerful n-grams technique [1,2].

The other approach is based on denotating information. Every relevant piece of information is augmented with data describing the information object, so-called: metadata. Metadata can be seen as filters as they distribute information according to classes, such as a name, an address, a keyword, etc. Looking for the name of the person Watt, we only have to look in the class of authors, whilst looking for the notion Watt (as a measure for electric power) we only have to look in the class of keywords belonging to the field of electric engineering. Due to the ambiguity of words, normally metadata are added by hand or based on the structure of the information object, e.g., a document. In a standardised environment we can infer with 100% certainty what the name of the author is, which is impossible if we deal with a document with an arbitrary structure in a language we don't master.

It goes without saying that both approaches, purely statistical and pre-coordination are needed in a real life environment. Statistical approaches have a number of obvious problems (lack of semantic knowledge, inability to interpret irony or casual references), while full pre-coding by the author might on the one hand be impossible to achieve, and on the other hand prevent the browsing reader to stumble on unexpected relationships or cross-disciplinary similarities. The challenge is how we can prepare information in order to enable quick and relevant retrieval, while not overburdening the author or indexer.

In adding metadata to documents, more and more computer assisted techniques are used. Some types of metadata are more or less obvious, e.g., bibliographic information, while others demand a deep knowledge of the content at issue. At the content level we deal with authors who are the only ones who can tell us what they want to convey and professional indexers who try, with the help of systematic keyword systems, to contextualise the document into a specific domain. In particular the last craft is creating essential added value by securing idiosyncratic individual documents into a domain context, by using well designed metadata systems in the form of thesauri and other controlled keyword systems.

We are currently working on the design of a system which enables the author to add as much relevant information as possible to her/his work in order to enhance retrievability. As writing cultures do change as a

result of the technology used, we propose to fully exploit the electronic capabilities to change the culture of authoring information. In such an approach, it is the author who contextualises the information in such a way that most ambiguities are pre-empted before release of the work. Such an environment is much more demanding for the author and editor, but ensures that the context of the work is well-grounded.

To build a useful development environment, in this contribution we define different categories of metadata, that are created, validated and used in different stages of the publishing process. Given the importance of metadata, we believe it should be treated with the reverence usually reserved for regular data, in other words, we need to worry about its creation, standardisation, validation and property rights. In this contribution, we want to explore how metadata is used, and consider the issues of versioning, standardisation and property rights. We then come up with a proposed, and very preliminary, classification of metadata items, and discuss some issues concerning the items mentioned. As we believe that metadata should be treated on equal footing as the objects they describe, in other words metadata are information objects in themselves, we show that all issues that pertain to information objects also pertain to metadata.

This contribution is meant to support our own work in building an authoring environment, and therefore does not present any conclusions yet- but we invite responses to this proposed classification and the issues at hand (versioning, validation, standardisation and property rights of metadata). Preferably, based on comparison of documents of different scientific domains, as it turns out that different domains can have substantial differences in structure and style. As is clear from the above and in particular from the table, many issues are still uncertain and in full development. For the design of an easy to use and versatile author environment, where the author can quickly denote her/his own writing and create and name the links to connotate the work, an analytically sound scaffolding is needed before such a system can be built.

Below we discuss a classification of metadata leading to an overview presented in a table. Items in the table refer to further elaboration via hyperlinks. As this presentation also has to be printed, in this version the elaborations and digressions are located linearly as sections after the table.

2. Classification of metadata

2.1 Different types of Metadata

In first approximation we make a distinction into three broad categories of metadata, which are accompanied by three uses of information:

- Type: Descriptive of content. Here we deal with typifying information that pertains intellectual knowledge needed for understanding and placing the work in context. Typical items are: the author's name, keywords & classification codes, an abstract, captions to various enhancements such as figures. etc. It can be argued that author's names, abstracts, captions and references are not metadata, but simply content. However, data about data are not necessarily atomic. An abstract denotes a story, hence, we have included it in the table.
  Use: Interpret and validate. As the reader normally is disjoint in time and place from the originator the interpretation of a work depends on its context. Note that this context is not only a matter of proper semantic indexing, but also defined by the technology used. If the original is handwritten on parchment or typed with 8 bit WordStar, the reader can make interferences about the cultural/technological state-of -the-art at the time of creation.
- Type: Descriptive of location. In this category we deal with traditional bibliographic references and their modern extensions such as the Digital Object Identifier (DOI) as well as status information such as draft (normally on a home page), preprint (on a home page and on a preprint server), revised version, final version (in a certified journal from a publishing institution), etc. In web-based publishing many versions of the same article abound, and knowledge of an object's location has to be augmented with knowing its status. Location thus means physical location as well as location in the added value chain from draft to certified document.
  Use: 1-Locate and connect. Here we deal with the traffic to and from data such as links to a work, to an author/subject index, or between works.
  2- Interpret and validate. If it is located on a preprint server, it can receive a different scientific status than if it is located on an online version of a high-impact journal.
- Type: Descriptive of format. In an electronic environment we are blessed with a plethora of technical rendering possibilities. Hence, every information object needs a complete description of its technical format, so in this category we deal with issues such as: presentation versions (txt, pdf, wrd, wpd, html,

etc., etc.) and in structured environments with descriptors such a Document Type Definition (DTD) and XML data standards (SVG, MathML, etc.).
Use: Manipulate. This can involve e.g. rendering certain data types or running programs. We have to know how to represent the information or how to use the metadata for statistical approaches or datamining.

2.2 Creation

Metadata can be created by different parties - authors, editors, indexers and publishers, to name a few. It is important to realise that at some times, the creating party is not the validator; also, if the creating party is not part of the versioning cycle, the party creating the latest version can be not aware of necessary updates in the metadata. Therefore, only the creator can add and validate such items as her/his own name or references to other works. Additional metadata can be generated by machine intervention, such as automatic file-type and size identification, whilst professional indexers, be it by hand or computer assisted, will add domain dependent context to a work.

2.3 Validation

Very often, metadata is not validated per se. For convenience's sake, it is often assumed that links, figure captions, titles, references and keywords are correct. An extra challenge in electronic publishing is the validation of non-text items - for one thing, most reviewers and editors still work form paper, thereby missing are hypertextual and/or interactive aspects of a paper (hyperlinks that no longer work are an obvious example of this problem).

2.4 Rights

The role of Intellectual Property Rights (IPR) and Copyright in particular, is a hot issue in the discussions on so-called self-publishing. A great deal of difficulty is in the differences between the various IPR systems, in particular between (continental) Europe and the US.
However, besides this issue, electronic publishing generates a series of even more complicated questions that have to be addressed. As metadata allow the retrieval of information, they become "objects of trade" by themselves. Below we only indicate some issues pertaining to our discussion. A more detailed overview on the complicted legal aspects in ICT based research is given in Kampermann et. al ([3] and references therein). This short list below, shows that the heated debate on the so-called copyright transfer (or really: reproduction rights) from the author to a publishers is only a small part of the issue. Metadata as information objects face at least the same right problems as the document per se.

- What is a work? The role of databases
  An electronic document is a well-defined set of different kinds of elements: texts, images, tables, sets of hyperlinks, & (meta)data. The E-document is an envelope of independent objects. The E-document can then be considered as a new object with various levels of granularity that can be accessed separately. By nature an E-publication is part of a virtual world-wide database. As soon as works are loaded on a (publisher's) web-site and value is added by, e.g.: the maintenance of links, the conversion to a standardized storage scheme etc., we can speak of a database that can claim the database protections given in the EC council Database Directive [4]. Here, in Article 1, a database is defined as a collection of independent works, data or other materials arranged in a systematic or methodical way and individually accessible by electronic or other means. A publisher's database becomes an integrated whole, an object by itself and can claim rights! Hence, the fact that the database is ruled by metadata has a crucial impact on the IPR's of authors.
- Metadata structures as object
  In order to fully exploit the electronic capabilities an author has to create his/her work according to well-defined rules that enable storage in a multi-media format. The author creates a work including a metadata structure that guides the reader. The presentation and the content in the electronic version are converging to one representation. In a controlled publishing environment a publishing organisation (commercial or not) creates and maintains the metadata structure. This means that at least the intellectual ownership is with that organisation and the added value to the "database" in which this structure is implemented is a genuine new creation. In case of the Open Archive self-publishing initiative it is the author who adds a limited set of meta-data his/herself.
- Controlled keyword systems and thesauri.

The design and maintenance of a thesaurus or ontology is intellectual labour and hence can be considered a work with its own IPRs. Different parties in the value chain can generate profit from this. In a world where documents (in one version or another) swarm around in cyberspace, the keys to disclosure become an obvious commodity.

- Form versus content (content driven publications)
  The whole new industry of SGML/XML declarative languages is geared towards presentation independent storage and the capability to "render" the "content" on different "platforms". So called: single source, multiple delivery publishing. The present IPRs deal only with the "content".
- Real multimedia "documents" (layout driven publications)
  A scientific publication is a mixture of text and non-text elements and in some case even non-text elements only. This original version of a scholarly publication will be a multi-media "document". The paper instantiation becomes a spin-off, needed for those who want to carefully read and annotate the work. But this version is not necessarily (and in the near future even pertinently not) the e-version minus the "unprintable" objects. A text for reading demands another grammar than a multi-media document. Hence, a scientific publication will consist of two or more presentation forms, that all need certification, authentication and validation, just like the old paper-only publication. Each form presents an independent entity, and deserve independent IPRs.
  The publication environment becomes intrinsically a collection with added value, due to the structuring and interlinking of the elements. New extra value can always be added by keyword and classification indexes as well as link taxonomies (different kind of hyperlinks, each with a meaning of why the linking is added). Those extra metadata systems are creative products with their own IPR.
  There is a difference between a real multimedia document, where the various expressions (text and non-text) are a united whole and the present-day patch works of various types of media (in fact multiple-media documents). Real hypermedia documents (an integration of hypertext and multi-media in which time-lines and spatial lay-out are well-defined) will directly be specified in terms of the final presentation (lay-out driven), the segregation between structure and presentation disappears. In such a case Database directive art.5.b, that allows database owners to carry out.....translation, adaptation, arrangement and any other alteration, becomes under heavy pressure, as form and content together establish a creative whole.
- Software
  Apart from these issues, we also have to consider that e-publishing requires a series of software licences from the Operating System to the Video Manipulation Package. All those rights become an intrinsic element of the publication and readers need to know which licences they need, even for simply reading a text. Hence, an extra system of metadata describing the required software and its parameters (version, single or multi-user, etc.) is appearing.

2.5 Metadata classification

Using the categories defined above, we can come to a first list of metadata items, that include comments on their usage, creation/validation and rights, and define a number of issues, that are described in the paragraphs below.

What is it	Category	Who creates	Who validates	Who has rights	Issues
Author name	content	Author	Author	Author	Unique author ID (see below 3.1)
Author affiliation	content	Author's Institute	Editor? Publisher?	Author?	Corresponding author address only? / Present address vs. at the time of writing. In other words is the article coupled to the author and her institution during creation, or does an article follows an author in time.
Author index	content	Publisher	Publisher	Publisher/Library	Author name issues (Y. Li issue, see below 3.1)

Keywords	content	Author, editor, publisher, A&I service, library, on-the-fly	Editor, publisher, A&I, library	See section 2.4	Multi-thesaurus indexing (see below 3.2) Ontologies
Abstract	content	Author, A&I service	Editor, A&I editor	Author/A&I service	Types of abstracts? Usage of abstracts? (see below 3.3)
References	location	Author	Editor, Publisher	None for individual reference; document collection - yes	DOI, http as reference; link reference to referring part of document; versioning! See also Links (below 3.4)
Title, Section division, headers	content	Author/publisher	Publisher	Publisher?	Presently based on essayistic narratives produced for paper
Bibliographic info (publisher's data)	location	Publisher	Publisher	Publisher (TM)™	DOI refers to a document, but is intrinsically able to refer to a sub-document unit. No pagination in an electronic file, referencing is now point-to-point instead of from page-to-page.
Bibliographic info (Other data)	locate	Library	Library	Library	Multiple copies in a library system, signature, etc. Does this all evaporate with the new license agreements, where the document is hosted at the Publisher's database?
Clinical data	content	Author	Editorial	Doctor/patient?	Privacy; standardisation; usage?
Link (object to dataset, object to object)	location/ content	Author Publisher	Publisher	Author? Publisher?	Are information objects (see below 3.4)
Multimedia objects Visuals, Audio, Video, Simul-(Anim)ations	content/ format	Author, Publisher	Editor? Publisher?	Rights to format (cf. ISO and JPEG) vs. rights to content	Who owns SwissProt nr? Genbank ® nr? Chemical structure format? JPEG org? Issue of layout-driven vs. content-driven data Who validates the scientific value of such an object? We don't have yet referee standards as we have for text.
Document status, version	content	Editor, publisher, (author for preprint/OAI)	Publisher	Publisher	Version issue (see below 3.6) Updated version in preprint/Open Archive Initiative (OAi)I - which is the original? Multiple copy problem; virtual journals
Peer review data	content	Reviewer	Editor	Reviewer?	How to ensure connection to article? Privacy? vs Versions of articles? Open or closed refereeing procedures
Document	content/ location/ format	Author, Publisher Reviewer	Editor, Publisher	Author ("creator")	Integrity of components that make up document; Versioning. Intellectual ownership vs reproduction rights (see also 2.4)

DTD	content/ format	Publisher	Publisher	Open source, copyleft?	Versioning? Standard-DTD (see below 3.7) (Dublin Core)? Ownership
Exchange protocols e.g. OAI	locate/ format	Library, Publisher, archive	"Creator"	?!	Rights! Open standards Original copy issue standardization (ZING);
Document collection - Journal (e.g. NTvG)	content/ location/ format	Editor/Publisher	Editor /Publisher	Publisher	Integrity of collection - multiple collections E-version versus P-version
Document collection - Database (e.g. SwissProt)	content/ location/ format	Publisher - Editor?	Publisher	Organization?	Validation? Rights?
Data sets collaboratories - Earth System Grid	content/ location/ format	Federated partners	Nobody!	Creator?	Validation? Usage?

3. Some issues

The following elaborates on some of the issues raised in the table in the previous paragraph (connected by hyperlinks in the online version). This elaboration is needed as only after a full understanding of the qualities and values of the various types of metadata and their mutual relationships we can start with the system requirements of new types of authors' environments to be designed in close connection with the storage and retrieval environment of genuine electronic -multimedia- documents.

3.1 Unique Author ID

The demand for an unique author id is as simple as reasonable. However, in the real world we encounter the following caveats:

- How do we know that it is the same author? Many people have the same name such as: Y. Li, T. Sasaki, Kim Park, or Jan Visser.
- Many transliterations from non- European languages into the Latin alphabet are different. Happily most academic library systems now do have concordance systems in place. But still, many uncertainties remain in cases such as Wei Wang (or Wang Wei).
- Authors change address, institutes change name (and address), and this amplifies the problem.
- Authors sometimes change their name after marriage, immigration or change of sex. This might be minor problem to the above mentioned, but is a persistent and frequently occurring problem .

So, do we want to use a social security (or in The Netherlands SOFI) number or picture of an iris scan? Or even introduce a Personal Publishing Identification Number (PPIN)?
A lot of practical and legal issues still stand in the way of true unique identification, but first steps are being set on this path by publishers, agents and online parties to come to a common unique ID - the INTERPARTY initiative being one of them.

3.2 Controlled Keyword Systems

Indexing systems are as old as science. The ultimate goal is to assign an unambiguous term to a complex phenomena or reasoning. As soon a something has a name, we can manipulate, use and re-use the term without long descriptions. In principle, a numerical approach would be easiest, because we can assign an infinite number of ids to an infinite number of objects. In reality, as nobody things in numerical strings, simples names are used. However, as soon as we use names we introduce ambiguities as a name normally has multiple meanings
A known problem is that author added keywords normally are inferior to keywords added by trained publishing staff, as professional indexers add wider context where individual authors target mainly on terms

that are fashionable in the discussion at the time of writing, as the experience in the journal making industry learns. Adding uncontrolled index terms to information objects therefore rarely adds true descriptive value to an article, a prime reason to use well-grounded thesauri and ontologies.

A so-called Ontology is meant to be a structured keyword system with inference rules and mutual relationships beyond "broader/narrower" terms. At present we are still dealing with an mixed approach of numerical systems such as: Classification codes, e.g. in chemistry or pharmacology, and domain specific thesauri or structured keyword system such as EMTREE and MeSH terms in the biomedical field. Therefore, most ontologies still rely on existing indices, and ontology mapping is still a matter of much debate and research. Currently, multifarious index systems are still needed, based on the notion that readers can come from different angles and not necessarily via the front door of the well established Journal Title. Index systems must overlap fan-wise and links have to indicate what kind of relationship they encode. The important issue of rules and particular the argumentational structure of these roles is part of our research programme and discussed elsewhere [5, 9].

3.3 Abstracts

The history of abstracts follows the history of the scientific paper. No abstracts were needed when the number of articles in a field was fairly small. Only after the explosion of scientific information after WWII we see the emergence of abstracts as a regular component of a publication. Abstracting services came into existence and in most cases specialists wrote abstracts for specialised abstracting journals (like the Excerpta Medica series). Only after the emergence of bibliographic databases the abstract became compulsory as it was not yet possible to deliver the full text. After a keyword search, the next step towards assessing the value of retrieved document identifiers was by reading the on-line abstract. In an electronic environment (where the full article is as quickly on the screen as the abstract) the role of the abstract as an information object is under scrutiny, since for many readers, it often replaces the full text of the article. As already said in section 2.1, abstracts are identifiers for a larger information object: the document. In that sense an abstract is a metadata element.

In a study at the University of Amsterdam [6] to assess the roles of the abstract in an electronic environment, the following distinctions are made :

Functions:

1. Selection. You cannot read all articles published. Facilitates choice.
2. Substitution. All relevant information is in the abstract, e.g. essential experimental results.
3. Retrieval. "In fact, the ideal abstract from an indexer's point of view is a string of keywords linked into an easily read sentence" .
4. Orientation. In supporting those who read (parts of) the source text. In an electronic environment it can be the linchpin of all components.

Type of abstracts:

1. Characterizing. A brief indication of what is it all about. Often a clarification of the title. Often author created.
2. Slanted. Oriented to a well-defined audience. E.g. abstracts of biological articles for chemists. Often made by A&I service.
3. Extensive. Useful if the source text is not easily available. Often made by editor/ domain expert.
4. Balanced. Reflects all phases of the reasoning. Imported if the abstract fulfills and orientation function.

This analysis shows that the database field "abstract" now has to be endowed with extra specifying denotation. As our research is on design models for e-publishing environments, we have to realise that at the authoring stage of an abstract a clear statement about function and role is needed, as more abstracts -of a different type- might be needed to cater for different reader communities.

3.4 Hyperlinks

As already discussed above, analysing components of a creative work into coherent information objects, means that we also have to define how we synthesize the elements again into a well behaving (new) piece of

work. The glue for this puzzle are the Hyperlinks. An important aspect in our reserach programme is to combine denotative systems with named link-structures that add connotation to the object descriptors. By integrating a proper linking system with a clear domain-dependent keyword system, a proper context can be generated.

If we analyse hyperlinks we have to accept that they are much richer objects than just a connection sign, as:

- Somebody made a conscious decision to make that link and so a link has formally an originators or Author.
- A link has been made during a particular process where the relevance of making the link became clear. E.g., in a research process it becomes clear that there is a relationship with other research projects. Hence, a link has a creation date. In an even more fundamental approach one can say that the creative moment of linking one piece of information to another is a discovery and is linked to a creator like any other invention or original idea.
- Hyperlinks belong to a certain scientific domain. A reference in geology will normally not point to string theory. Hence, the point where the link starts is an indication for the information connected. It goes without saying that this is never completely the case as a geologist might metaphorically link to a publication on The beginning of Time according to mathematical physics.
- Links can carry information on the reason of linking (see below).
- Most important, links carry knowledge! They tell us something about relationships in scientific discourse.

All in all, hyperlinks are information objects with creation date, authorship, etc. and hence, can be treated like any other information object. This means that we have to extent our discussion of metadata as data describing information to hyperlinks.

Apart from the obvious attributes such as author, date, etc. we can think about an ontology for links. This ontology will be on a more abstract level than an ontology of objects in a particular scientific field as we here we deal with relationships that are to a large extent domain independent.
A first approach towards such system might go as follows:

A) Organisational

- Vertical. This type of links follow the analytical path of reasoning the relations are e.g., hierarchical, part-of, is a, etc.
- Horizontal. This type of link points to sameness and look alikes, such as: see also, siblings, synonyms, etc.

B) Representational

- The same knowledge can be presented in different representations depending on the reading device (PDA, CRT, Print on paper) or by the fact that some information can be better explained in an image, a table or a spread- sheet. Therefore we have a family of Links that relate underlying information (or data-sets) to a graph, a 3D model, an animation or simply a different style-sheet. As these links will also related different presentations of the same information, if available, many of these links might be generated automatically.

C) Discourse
The great challenge in designing a link ontology, and metadata system is in developing a concise but coherent set of coordinates. As discussed in more detail elsewhere [7, 8].
We suggest the following main categories:

- Clarification (link to educational text)
- Proof (link to mathematical digression elsewhere, link to law article, etc.)
- Argument e.g., for/ against different author

In conclusion: as links are information objects we have to be aware of validation and versioning the SAME way as textual or visual objects and data-sets!

3.5 Standardization

In an electronic environment where documents (or parts thereof) are interlinked, no stand-alone (piece of) work is created/edited/published anymore. All creative actions are part of a network. So, all parties need to discuss and use standards: (partly) across fields, (certainly) across value chains. However "The great thing about standards is that there are so many to choose from..." and that they evolve all the time.

In library systems, we rely on a more or less certified system of index terms such as Machine-Readable Cataloging (MARC) records, where a distinction is made between: Bibliographic, Authority, Holdings, Classification and Community information. In a more general perspective we see all kind of standardisation attempts to ensure interchange of information in such a way that the meaning of the information object remains intelligible in the numerous exchanges over Internet. {See e.g.. The National Information Standards Organization (NISO) in the USA for the Information Interchange Format and The Dublin Core Metadata Element Set}.

An immediate concern is the level of penetration of a standard in the field and its, public or commercial, ownership. Who has the right to change a standard, who has the duty to maintain a standard, how is the standardisation work financed and who is able to make financial gains out of a standard? For that reason the discussion of standardisation and Open Standards in particular are crucial in this period of time.

3.6 Versioning and modularity

Today's authoring tools allow the easy production of many different versions of a publication prior to certification. Often drafts are sent around to colleagues for comments. It is not unusual that drafts are hosted on a computer system that allows others to approach the directory where the draft is located. Comments are often written into the original work and returned with a new file name. That way, many different versions of a document float around without any control and without any guarantee that after the drafting process is closed an a final work is born, all older versions are discarded. The same problems occurs again in a refereeing process if the paper resides on a pre-print server. All this forces the installment of a clear versioning policy. In practice this means that the metadata describing a work (or parts thereof) must have unambiguous data and versioning fields, indication not only the "age" of the information but also its status.

An interesting new phenomenon appears here. As is well known, in may fields so-called salami publishing is popular. Firstly a paper is presented as short contribution on a conference, than a larger version is presented on another conference and after some iterations, a publication is published in a journal. It is also common practice that people publish partial results in different presentations and then review them again in a more comprehensive publication. This practice can be overcome if we realise that an electronic environment is essentially defined as an environment of multiple and re-use of information. The answer to the great variety of versions and sub-optimal publications might lie in a break up of the linear document into a series of inter-connected well defined modules. In a modular environment the dynamic patchwork of modules allows for a creative re-use of information in such away that the integrity of the composing modules remain secured and a better understanding of what is old and what is new can be reached. Such an development is only possible if the description of the various information objects (or modules) is unique and standardised [7, 8, 9,10].

3.7 DTD

As said in section 2.1, the description of the format of the information is an essential feature for rendering, manipulating and datamining the information. This means that we need a full set of technical describers identifying the technical formats as well as identifiers that describe to structure of the shape of the document. Opposite to the simple technical metadata, e.g., are we dealing with ASCII or Unicode, the metadata that describe the various linguistic components and the structure of a document are interconnected one way or the other. This means that we need a description of this interconnection, hence metadata on a higher level.
A Document Type Definition (or its cousin, a Schema) defines the interrelationship between the various components of a document. It provides rules that enable checking (parsing) of files. For that reason a DTD, like an abstract belongs to the metadata of a document.

- E.g. A name field MUST have a Family name, must have at least a first initial, may have a second name/initial, may have pre- and post particles.

Based on such a skeleton DTDs and Style Sheets can be designed that keep the integrity of the information

(up to a -to be defined- level) tailored to various output/ presentation devices (CRT, handheld, paper, etc.).

- E.g. If a full first (or subsequent) name(s) is available, then on paper it is spelled out in its entirety, but on a handheld we only see the first initial.
- E.g. If colour is essential but the output device does not support colour, a message is added to the presentation.

Within this problem area it is important to mention the difference between content driven publications, i.e. publications that allow different presentations of the same information content and can be well catered for by a DTD and lay-out driven publications, which are publications where e.g., the time correlation between the various elements is essential for the presentation. See e.g. the work done at the CWI [11].

*) Also at : Van der Waals-Zeeman Institute, University of Amsterdam and the Research in Semantic Scholarly Publishing project of the University Library, Erasmus University, Rotterdam

4. References

1. Marc Damashek. Gauging Similarity with n-Grams: Language-Independent categorization of text. Science. vol 267. 10 February 1995. pp 843-848.
2. Alexander M. Robertson and Peter Willett. Applications of n-grams in textual information systems. Jnl of Documentation vol 54. no.1 January 1998, pp 48-69.
3. European Research Area Expert Group Report on: Strategic Use and Adaptation of Intellectual Property Rights Systems in Information and Communications Technologies-based Research Prepared by the Rapporteur Anselm Kamperman Sanders in conjunction with the chairman Ove Granstrand and John Adams, Knut Blind, Jos Dumortier, Rishab Ghosh, Bastiaan De Laat, Joost Kircz, Varpu Lindroos, Anne De Moor. EUR 20734 — Luxembourg: Office for Official Publications of the European Communities. March 2003 — x, 78 pp. — 21,0 x 29,7 cm. ISBN 92-894-6001-6.
4. Directive 96/6/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases. Official Journal L 077, 27//03/1996 p.0020-028.
5. Joost G. Kircz. Rhetorical structure of scientific articles: the case for argumentational analysis in information retrieval. Jnl. of Documentation, vol.47, no.4, December 1991, pp. 354-372.
6. Maarten van der Tol. Abstracts as orientation tools in a modular electronic environment. Document Design, vol. 2:1, pp.76-88, 2001.
7. J.G. Kircz and F.A.P. Harmsze. Modular scenarios in the electronic age. Conferentie informatiewetenschap 2000. Doelen, Rotterdam 5 april 2000. In: P. van der Vet en P. de Bra (eds.) CS-Report 00-20. Proceedings Conferentie Informatiewetenschap 2000. De Doelen Utrecht (sic), 5 april 2000. pp. 31-43. and references therein.
8. Joost G. Kircz. New practices for electronic publishing 1: Will the scientific paper keep its form. Learned Publishing. Volume 14. Number 4, October 2001. pp. 265-272.
   Joost G. Kircz. New practices for electronic publishing 2: New forms of the scientific paper. Learned Publishing. Volume 15. Number 1, January 2002. pp. 27-32. See: www.learned-publishing.org
9. F.A.P. Harmsze, M..C. van der Tol and J.G. Kircz. A modular structure for electronic scientific articles. Conferentie Informatiewetenschap 1999. Centrum voor Wiskunde en Informatica, Amsterdam, 12 november 1999. In: P. de Bra and L. Hardman (eds). Computing Science Reports. Dept. of Mathematics and Computing Science. Technische Universiteit Eindhoven. Report 99-20. pp. 2-9.
10. Frédérique Harmsze. PhD Thesis, Amsterdam, February 9, 2000. A modular structure for scientific articles in an electronic environment.
11. Jacco van Ossenbruggen. PhD Thesis, Amsterdam, April 10, 2001 . Processing Structured Hypermedia- A matter of style.

URL's mentioned

DOI: http://www.doi.org
Elsevier: http://www.elsevier.com
Dublin Core: http://dublincore.org/
Earth Systems Grid: http://www.earthsystemgrid.org/
Genbank: http://www.ncbi.nlm.nih.gov/Genbank/index.html
Interparty: http://www.interparty.org/
JPEG-Org: http://www.jpeg.org/

JPEG: http://www.theregus.com/content/4/25711.html
KRA: http://www.kra.nl
Marc: http://www.loc.gov/marc/
NISO: http://www.niso.org/standards/
NTvG: http://www.ntvg.nl/
OAi: http://www.openarchives.org/
Ontologies: http://protege.stanford.edu/ontologies/ontologies.html
Research in Semantic Scholarly Publishing project: http://rssp.org/
Swissprot: http://www.ebi.ac.uk/swissprot/index.html
Trec: http://trec.nist.gov/
ZING: http://www.loc.gov/z3950/agency/zing/zing-home.html