XML/SGML TECHNOLOGY IN THE FIELD OF MEDIATING CULTURAL HERITAGE RECORDING - ARCHIVING - INFORMATION RETRIEVAL IN ORDER TO MEDIATE CULTURAL HERITAGE[1]

DR. JOHANNES PALME, BERLIN

Dr. Johannes Palme
Institut für Terminologie und angewandte Wissensforschung
(Institute for Terminology and Applied Knowledge Research)
Am Köllnischen Park 6/7, 10179 Berlin
Tel.: +49 (0) 30 86 20 88, Fax +49 (0) 30 86 20 87
E-mail jpalme@rz.hu-berlin.de

Recording, archiving and retrieving information are „the" tasks carried out traditionnally by archives, libraries, museums, by information and documentation centres. The focal point of activities differs in the several institutions:

- Museums collect, record and present knowledge. They are mainly dedicated to the aspect of presentation. The exhibition, i.e. the choice of objects, the style of presentation, the program in the context of the exhibition enable the visitor to participate in the exhibition. The museum can become a meeting place as well as a place of discussion or a place of learning.

- The tasks of archives are mostly defined by law. The law defines which kind of material has to be archived as well as the space of time in order to secure the heritage. Predominant is the providence principle. Additionally selection criterias have to be defined as well the availibility for long-term use to be guaranteed.

- Libraries own unique historical collections, too, and have taken over archival functions in this sense. Libraries have focussed their activities to the field of cataloguing and making information available for research, for teaching and training. The parts of libraries, available directly to users, are usually presented in a systematic manner. Information retrieval through catalogues is possible via formal criteria as well as through subject headings and classifications.

- Documentation focusses generally to a special subject and indexes it in a extensive way (by the use of thesauri, terminology pp.) Beside bibliographical references and catalogues, which are managed by libraries and archives as well, data and facts are indexed, classified and made available.

- Scientific edition aims to prepare important works in the field of art and humanities for research, managing the texts as historical documents. Historic-critical editions meet the highest requirements by presenting a text together with the history, the delivery and the receiption of a work. Beside these criteria annotations and text variations are integrated.

All these activities have in common that they record, archive and organize the cultural heritage (of a city, of a region, of a nation, about a subject).

[1] The talk held at EVA'99 included the presentation of the CD-ROM „Kulturerbe digital". The CD-ROM is available via the German Library Institute, situated at Berlin. A email subscription form is available via the homepage of the Institute of Terminology and Applied Knowledge Research (http://www.itaw.hu-berlin.de). In order to understand the full opportunities of SGML-based publishing and visualisation the reader of the article should install the CD-ROM and explore the presentation and the retrieval facilities.

1. The concept

The independent project BIADOK_Publikation (supported by the Institute of Terminology and Applied Knowledge Research, situated at Berlin) was founded in 1996 and works in the field of XML/SGML based knowledge structuring and browser visualisation. The description and presentation of cultural objects implies the use of different kind of media: books, press clippings (articles in newspapers and journals), audiovisual material (images, films, sound documents), manuscripts, objects and facts have to be collected, analysed and presented. This manner of managing material is a vivid example for the integration of multimedia elements and the presentation of complex, multivarious linked knowledge structures. Computer based information systems in the field of culture and sciences (for example of a museum or a research institute) have some advantages in comparision with conventional ways of scientific research and publishing. It applies to the rationalization of working procedures by the online access to the permanently updated database, the shared use of information (facts, texts, pictures). Furthermore a computer based information system allows a differenciated access to the material and an improved use of the information. Such an information system offers improved services to the users (access to information by databases, profile services, pp.) and supports the marketing opportunities of the institution itself, for example to raise funds, to make the institution well known.

1.1. SGML - Standard Generalized Markup Language

The knowledge structuring bases on the ISO standard 8879 SGML (Standard Gerneralized Markup Langauge). SGML has been developed to manage, maintain and exchange complex structured technical information.

- The structue contains the logical organisation of a document. A document consists always of different logical parts. A book, for example, consists of chapters with titles, sections and subsections. A section itself consists (in any sequence) of other elements, for example lists, tables, graphics or simply text. All these parts or elements of the documents have a certain hierarchical relationship. The sections of a book are generally to be found in a chapter and therefor one hierarchical level below the chapter.
- The content of the document (document instance) comprises the text itself and - if existing - non-textual elements, e.g. graphics, images, multimedia elements, for example video tapes, sound documents can be integrated into such a document.
- The layout (style) defines the way of presentation, the way of visualisation on paper, on the screen, on CD-ROM. Style editing means the linking of structure elements to layout elements: which fonts is used for titles, lists, texts, indicates emphasis pp. ? The layout has to be separated strictly from the logical structure, it only concerns the way of presentation with regard to the readability on screen or on paper. The layout does not belong to the SGML application in a narrower sense.

A SGML application generally consists of two parts, stored together in one or separately in two or more ASCII coded files: the document type defintion (DTD) on one hand and the SGML marked document instance on the other. The document instance is characterized as marked, because the document contains beside the text „tags", put in the text at the beginning and the end of the separate elements.

The advantages of SGML result from the features of SGML marked documents: the separation of content and context allow the platform-, software- and application independent management of data and layout. The validity, i.e. the correct element use, of the documents can be easily proved by computerbased parsing the document instance and the document type definition. SGML is software independant. That means, it is always possible to change the tools for publishing and visualisation without any compability problems. Another advantage is the fact, that the document instance only contains structural and no layout information. This means, that different styles can be linked to the same document with regard to different purposes. This is important, if documents have to be published on different media. SGML structutred documents can be easily adapted to further developments in the field of word processing and desk top publishing, because only the style of presentation has to be changed. The expediture for the data management is quite low: elements and elements groups can be changed or exchanged automatically, they can be deleted automatically. Restructuring and reorganising of document and text structures can be carried out quite easily on the basis of the DTD, too.

The mostly known SGML application is HTML (Hypertext Markup Language). HTML ist one special document type definition, based on the SGML standard.

1.2 XML - Extensible Markup Language

Quite new is XML, a subset of SGML. XML aims to use the attributes of SGML for processing and publishing documents on the Internet. It has been developped as intermediate stage, simplifying the use of SGML and guaranteeing the interoperability to HTML. XML can be introduced by those, who would like to deliver information through the Internet and need to go beyond the opportunities of HTML. Potential applications are electronic books, financial transactions (e-commerce), technical documentation, chemical formulars, medical information, museum catalogues, encyclopedias pp.. The main differences in comparison to SGML are:

- the definition of a character set for the XML meta language (e.g. chinese, greek, latin based)
- no minimization, i.e. start- and end-tags have to be set and can not be minimized
- restrictions in the element declaration, for the attributes, within the use of entities in comparison to SGML (no floating elements)
- the extension of linking possibilities, based on the expriences within the implementation of HTML, HyTime and TEI
- XML documents are self describing, i.e. all needed information are to be found in the header of the XML document.
- unicode as a standard feature

The consideration for one approach (SGML, XML, HTML) depends on the requirements of the planned application.

1.3 The approach

In order to develop prototype publications the following step by step process has been proved as effective and practicable. The first step is the development of a document type defintion (DTD) by modifying existant publicly accessible DTDs or by making a new one with regard to the specific requirements of the SGML application. In the second step the source data have to be converted into a a valid SGML structured document instance. Texts, images, sound and videos were integrated into the application. Afterwards the knowledge structures has to be visualised for the use and retrieval through a SGML browser. The graphical interface and the retrieval facilities depend on the publication structures and the facilities of the browser used. Elements, to be put into the table of contents for hypertext navigation, have to be defined. Beside and above this, the typographical appearence of the elements on the screen, hyperlink facilities within the document and to external documents, different styles of view to the materials have to be defined. Certain elements, for example footnotes, copyright, tables, can be hidden and revealed via icons in order to enable a comfortable use and get a clearly structured view. The user friendliness and the use of SGML structures are improved by the use of search forms. The user is able to benefit of the SGML structures without knowing them by using the search forms.

2. The CD-ROM „cultural heritage - digital“

Prototypical publications, developed within the last three years as example applications of SGML in the field of archives, libraries, museums, documentation and edition, have been put together on the CD-ROM „Kulturerbe digital“. The examples of SGML based publications are presented through the DynaText 4.1 browser. The publications have been choosen in order to present different aspects of electronic publishing: the integration of multimedia elements (text, pictures and graphics, sound, film, 3-D-visualisation) and the fields of application. The CD-ROM tries to give an impression, which products can be developed on the basis of SGML in the field of documenting the arts and humanities. Two examples are presented during the talk, hold at EVA 1999.

2.1 Multimedia presentation in the museum

A find of the early middle age of Montcornet, Département Aisne / France, was documented. An iron belt-buckle, decorated with an animal in the style of the 7th century, is almost completely conserved. Until now only one half side was restored[2]. This publication focusses on the integration of external applications. The integrated graphics have to be seen as text explication via image and enable furthermore the launch of external applications. The used image material contrasts modern methods to traditional methods. Usually impossible things, for example the multiple look to the different steps of the working process, making transparent of the unique find can be carried out by using the mouse, clicking into the image in the electronic publication. Clicking, i.e. using the mouse, starts a film viewer or a program for 3-D-visualisation, enabling the change of the perspective of the view. A personal computer with CD-ROM device realises the visualisation, a pentium III chip with 450 Mhz is sufficient for a presentation without delay. This application can be seen as a first step towards the active participation of the visitor as well as a step beyond a „traditional“ presentation. The object can be turned freely around, it can be touched and enlarged.

2.2. SGML based film documentation

The film is a quite young medium. Since about hundred years stories and subjects are visualised by this way. Comparable to a play a text forms the basis, which is transfered into pictures (spoken text, action). In opposite to a theatre production a film can be copied and presented to spectators at different location at the same time. There is no direct interaction between the audience and the actors, because the film is a finished product. Film can not only be presented in the public (at the cinema), it can be seen by individuals (via television or video tape). In this case there is no (or only minimal) interaction between the spectators, implying another quality of perceiption as the perceiption in the cinema.

The centre of the film is the story. The story can be adapted of literary patterns (fiction, drama, tales), the adaption of historical events or completly fictional. The story of the „Hauptmann von Köpenick“, a film by Frank Beyer and Wolfgang Kohlhaase, was adapted from the play of Carl Zuckmayer. The play of Zuckmayer based on an article of a newspaper and characterizes the German society at the beginning of the 20th century in Prussia. The film, i.e. a video tape in the VHS format, has been digitized[3] (format: MPEG 1) and combined with the electronic version of the script. Additional information of the directors script has been integrated, for example the casting list and the list of roles. Differences between the script and the realised film were marked, too. With regard to the fact, that a prototype publication is presented, only a part of the film was digitized. With minimal expense publishing on different media can be realized (Internet, CD-ROM, DVD-ROM in the near future). Besides these advantages different forms of presentation, different views of the same materials can be made available, for example the view for the film-goer (spectator) as well as the view for film researchers or a multilingual approach.

[2] DResearch has supported us by giving the permission to integrate their application into our electronic publication.

[3] We thank the TZI of the University at Bremen for the digitization of the film with regard to our structural requirements.

Beside the aspect of archiving by digitizing other criteria have to be taken into consideration from the point of presenting films. Results of preparatory work and research to be carried out every time when someone builds up an archive or documents projects are made available together with the film and other belonging material. In doing so, existing information and film databases can be used as well as electronic newspaper archives. Furthermore new products can be developed. In addition to formal criteria (title, film director, film production firm, year of production, casting list, staff list, pp.) content oriented criteria (film subjects, themes) and technical criteria (camera positions, cutting techniques) can be marked up and made searchable. The parallel presentation of different versions of a film, the parallel presentation of comparable film subjects open new perspectives from the point of the methodology of film research as well as the perception and reception of film in general.

3. SGML based indexing

The extensive indexing of the material is the supposition to enable quick and comfortable information retrieval. Indexing enables information experts and users to find documents via different points of access: via formal aspects as the name of the author, title, subject, to get information about documents of the same author, subject - but different locations, to distinct different editions or different publication forms.

3.1. Indexing of electronic publications

In opposition to the way of indexing traditional media, which are physically available for catalgoguing, additional aspects have to be taken into consideration. In the case of electronic publication the direct transfer of bibliographical information is possible. Source herefore are the opening screen, the electronic publication itself, the accompaying material. Interfaces are needed in order to exchange the information between the parts of the information systems. Furthermore additional information is needed: the ways of access to the document, structured descriptive information to enable automated indexing. In order to guarantee the long-term use, archive information has to be stored. Archive information means the format of the original, the using and the access rights, the archiving format.

3.2 Concepts to index electronic publications

At this point it becomes evident that indexing should not be restricted to the field of structural object description (mainly alphabetical cataloguing) of electronic documents. Rather indexing has to look at the production process, because the retrieval facilities are established during the making of electronic publications. For example: the producer (author) of the pages in the WWW establishes and maintains the links, offers search forms, etc. How can electronic publications be indexed and made available in an effective way to users? One way to catalogue the electronic publication is to adapt existing data formats in the area of bibliographic description, for example by defining a field for electronic location and access. In the meantime the German MAB format (MAB = Maschine Readable Format for Libaries) was modified in this area on the basis of the American model. Another way is using the bibliographic information included in the SGML-based document structure to index the electronic publication.

To receive the relevant elements needed for alphabetical cataloguing, the DTD of the publication has to be analysed. The DTD contains the elements, their relationship (structures, hierarchies) and their frequency. Alphabetical cataloguing from the point of view of SGML is a special view to the electronic publication. This view includes only those elements, decided as being relevant for cataloguing (author, corporate source, title ...), comparable to the CIP record (Cataloguing in Publication) in printed books.

Developing this concept consequently farther on, new aspects have to be taken into consideration: SGML can be used for the management of object data. Librarian data formats and electronic library catalogues, finding aids, inventary lists pp. can be reflected in SGML by special DTDs. Catalogue data are 'lifetime' documents, too, intended to be archived and to be retrieved in the long term. SGML-based publications offer additional facilities. Beside well-known traditional structuring elements of printed publications (for example: table of contents, index, abstract, summary, footnote, annotation) the reader has further possibilities: hypertext structures to navigate within the publication, full-text search, search in the element structure (headings, figure captions, definitions, citations) as well as the possibility to follow links to external documents. Especially the search in sections of marked text, structured on the basis of the logical and hierarchical structures of SGML, has to be emphasised. The author determines the retrieval during the genesis of producing his document. With this comfortable search facilities the principle of subject headings in the context of subject indexing mislays significance. The principle of defined terms (descriptors) will supersede the principle of subject indexing. Here the working fields of library and documentation come together.

Besides the concept of „implicit" information, included in electronic documents, which have to be filtered and selected from the elements (structures), exists the concept of „explicit" information (meta data), delivered with the document. Examples for this concept are DublinCore, intended to index HTML publications on the Internet, the TEI[4] header and the Warwick Framework in order to integrate metadata into SGML-structured publications. Using this concepts of electronic publications, the transition of the relevant elements in conformity to the categories of the data formats and rules has to be carried out „only".

4. Recording - Archiving- Information retrieval in order to intermediate cultural heritage

Archives, libraries and museums collect and deliver the cultural heritage. Computer-based indexing and information retrieval enables a better access to the material and the improved usage. Organizing and structuring the knowledge is the presumption to intermediate cultural heritage. XML- / SGML structuring ensures the longterm and multiple use of structured information. The structures can be adapted without any problems to changing technological opportunities as well as to changing users needs.

An essential advantage of XML- / SGML structuring is the distribution on different media (printing, local, LAN, WAN - including WWW -) without much additional effort. Styles have to be devloped, based on the DTD in order to get a clearly arranged view. From the point of the user the retrieval facilities play the most important role: multiple retrieval facilities can be offered because of the underlying SGML structures (full-text, Bool'ean operators, truncation, SGML-structured search) through the browser with graphical interface on the local computer, on CD-ROM and via Internet. Search forms, supporting the user's search, can be integrated. Furthermor different points of view to the knowledge basis can be realized, for example for different levels of search, different access possibilities for different users groups pp.

Beside a pure SGML-based approach mixed forms of presentation can be put into practice. HTML forms can be used to perform searches in a database and to present the results in different ways. For example the results can be presented as a list in the HTML format. Content is lost by this way of visualisation. One step beyond is a XML-/SGML based presentation of search results via HTML search forms. Instead a simple list of results the user recieves a content based presentation.

We conclude: Mediation of cultural heritage is definend from the perspetive of knowledge structuring. The mediation comprises flexibility during the process of the creation of knowledge structures as well as flexibility in the process of creating retrieval facilites with regard to the information retrieval.

[4] TEI = Text Encoding Iniative, Initiative to encode and exchange digitized texts in the field of arts and humanities.