CONTENT DESIGN
METHODS AND TOOLS FOR THE CREATION OF PORTABLE HYPERMEDIA ARCHIVES

Notes for a proposed CID project, draft version 2, 1997-10-04

Donald Broady, Nada/CID, KTH

SUMMARY IN SWEDISH

Syfte: att utveckla nya metoder och verktyg för skapande av digitala arkiv varur användaren kan hämta material att införliva med sina egna tillämpningar.

Bakgrund: Utvecklingen av flyttbara informationsformat (HTML, SGML, XML, VRML, MIDI, Quick Time, etc) och plattformsoberoende protokoll (från TCP/IP till Z39.50) erbjuder potentiella möjligheter för användaren att kombinera innehåll (dvs text, bilder, ljud, video) från många olika källor i enlighet med sina egna behov. Exempel är en läsare som sammanställer sin egen personliga morgontidning från många olika nyhetsleverantörer, en lärare som ut källor på Internet eller CD-rom sammanställer ett skräddarsytt läromedel för sina elever, eller en grupp medborgare som utnyttjar EUs planerade webb-utgivning av medlemsländernas lagstiftning samt material från legal publishers för att sammanställa ett kompendium om ett visst lagstiftningsområde.

Behov: För att dessa möjligheter skall kunna realiseras krävs dels teknikutveckling, dels nya slag av kompetens. En delvis ny situation är att ”content designers” behöver verktyg och metoder som tillåter dem att skapa flyttbart innehåll, vars framtida användning de inte kan förutse och inte bör föregripa.

Genomförande: Återstår att resonera om.

Tänkbara intressenter: Skolverket, Handikappinstitutet…

1. BACKGROUND: PORTABLE HYPERMEDIA, PROMISES AND OBSTACLES

The development of portable information formats (HTML, SGML, XML, VRML, MIDI, Quick Time, GIS etc) opens possibilities for end users to combine content (text, graphics, sound, video, numeric information) from different sources in accordance with their own needs. You might wish to create your own personalised newspaper by combining material from different news providers. A teacher might wish to compile tailor made courseware for his or her class built on material extracted from repositories on the Internet or from CD-ROMs.

Such uses of digital resources are still rather rare, however. There are several reasons why the promises of integrated open hypermedia are not yet fulfilled.

One main obstacle is that most information providers prefer to market accomplished and polished products on for example a CD-ROM or on a web site rather than “raw material” that the users might adapt to their own needs. This is partly due to unsolved copyright and other legal problems as well as to commercial considerations. Today there exists no easy way to bill a user who extracts bits and pieces from various sources on different media (WWW, CD-ROM, local storage) or to remunerate the creators and information providers. Further, many system vendors and information providers are reluctant to deliver information in portable formats, for two reasons: in fear of improper duplication and in order to force the user to stay with their systems and their products. An unfortunate consequence is that albeit many publishing houses and other information providers use portable formats (e.g. SGML) in their in-house production systems, they ship products which content is stripped or confined to proprietary delivery technology. The content is expected to be used in a specific environment offered by a proprietary system or on a certain CD-ROM or web site. Users are doomed to run into problems if they wish to extract more complex content (including meta-data, link information, content markup, structural relations within and between files) in order to reuse it for purposes not foreseen by the vendors.

Other hindrances are due to the present state of technology. Even if the users happen to dispose qualified content in portable format, there is a lack of appropriate affordable systems and tools to manage such content. In theory, one could imagine a situation where the users build their own personal library for multi-purpose use by collecting information building blocks in for example SGML, HTML, XML, VRML or MIDI format. The development of international standards is mature enough to implement such applications. In large corporations and organisations you find specialists who to some extent work in this manner. There is, though, an obvious need for systems and tools that permit a wider public to manage complex content in portable standardised formats.

There is also in many fields a shortage of available high quality archives in portable formats, which in turn to some extent is due to lack of competence in the design of such archives.

2. CONTENT DESIGN

There is a need for a new type of professional competence in an emerging field that might be labelled content design.

It is often said that the “content industry” is getting the upper hand in the IT competition. Content itself is becoming a most valuable asset. To control the rights to text, images, music or movies might be just as important as to control the hardware and software used to manage the content or the telecommunication infrastructure and other publication channels used to distribute it. Publishing houses, movie and television companies, news agencies, image archives, and even the creators themselves (provided that they have not given up the rights to their means of production and their products) seem to be in a favourable position.

In accordance with this argument it might be useful to distinguish between three IT design domains: system design, presentation design, and content design. System design is a well-known domain. Presentation design is a much older domain which have been cultivated by book designers or theatre or music directors for many centuries. More recent sub-domains for presentation design are the creation of CD-ROM products or web sites. Finally, content design, i.e. preparation of the material that is to be presented, is an antediluvian domain. Oral storytelling had to be transformed when transferred to a scroll or a codex. Composers have provided content for the musical performers, authors and editors for book publication, and scriptwriters for the movie industry.

Content design as an IT domain in a narrower sense is, however, somewhat different. One of the most crucial differences is that content design might be separated from presentation design to a never before beheld degree. The potential benefits of this separation is not as yet fully realised. Still most book productions, music products, CD-ROMs or web sites are produced in the traditional manner, which means that the digital content is designed to fit a specific delivery medium.

However, as a consequence of the development of open systems and internationally standardised portable information formats there is a potential for new kinds of content design. From this perspective a content designer is someone who procures high quality raw material for unforeseen multi-purpose use. The content designer avoids preconceived conceptions on how the content is to be presented. He or she might create an information package including GIS conformant data to be used for drawing a map, for producing graphics, for calculations, for a 3D presentations of the topography of a geographic region, or for generating ID-attributes for cities or other localities treated in an interactive digital course-book in geography. He or she might produce MIDI-files to be used in bits and pieces by music students or performers. He or she might prepare a digital edition of an authors collected writings which in the future will form the basis for publication on paper, CD-ROM or Internet or perhaps for presentation as Braille, speech synthesis or large print for visually impaired.

In the publication industry a content designer is sometimes called a technical editor. Normally a technical editor as responsible for transferring material delivered by traditional editors into a specific product, today typically a CD-ROM or a web site. Thus, the technical editor might co-operate (often via editors) with authors, image creators, music performers and others who procure the input, as well as with specialists in user interface design, graphical design, visual arts etc in order to accomplish a desired result on for example a CD-ROM. All these kinds of competencies are needed also in content design in a more general sense, that is in a situation where the output format and delivery media is not settled in advance. A content designer has to co-operate with the same specialists as a technical editor who produces a closed end product.

In addition, an even wider range of competencies are needed in content design aiming at the production of portable hypermedia archives. On the one hand such archives have to be thoroughly structured so as to allow the user to navigate through the content and scan, search, filter, reorganise and reuse it for purposes that are not foreseen by the designer. To ensure that the users will be able to import the content into their own archives and systems, the formats (including meta-data formats, markup schemes and linking mechanisms) have to be portable and compliant with relevant standards. Proprietary and platform- and application-dependant indexing, keywording and linking mechanisms have to be avoided in order to allow content to be transferred to a wide range of environments and applications.

On the other hand, even if the content designer is to avoid precocious conceptions on how content will be used and presented, it is necessary to have a good grasp on available current information management and presentation technology, as well as a vision of the future. The content designer should be familiar with for example today’s presentation technology for delivery of content on paper, CD-ROM and on-line. In the most immediate future the content will probably be used either for local management or in end products on paper, CD-ROMs or web sites.

Though it is difficult to anticipate the future, there exists one golden rule: to pay attention to the development of international standards.

3. RELEVANT ONGOING STANDARDISATION

Currently one of the most interesting emerging standards for content management and delivery is XML (Extensible Markup Language), a language for advanced Web applications proposed by a working group of W3C (the World Wide Web Consortium). The most recent version (version 1.0, August 7th 1997) of XML is available at http://www.w3.org/TR/WD-xml-lang.html. XML is intended for the distribution on the Internet or intranets of content that is more complex than can be managed by the HTML-scheme. In short XML offers some of the advantages of SGML (for example user-defined tagging schemes, and the ability to represent hierarchical relations between elements) while renouncing the complexity of full-fledged SGML solutions.

Another important field for international standardisation concerns addressing and linking information. Current addressing mechanisms like HTTP seriously hamper the portability of content, for example since links break whenever the physical location of files is changed. The development of international standards for "public identifiers" etc. will hopefully result in a situation where digital resources might be addresses in much the same manner as we today use ISBN or ISSN to identify printed information. In order to refer to book you do not need to know its place on a certain shelf in a certain bookstore. You should not need to know the physical location of a digital content in order to address it.

A third important tendency in information standardisation is the change of focus away from "structured documents" (often identical to SGML files) towards information objects. If SGML were invented today it would be regarded as a meta-language not for document description but for information or content description.

A forth relevant field of standardisation is search and retrieval mechanisms. The standard "Information Retrieval (Z39.50): Application Service Definition and Protocol Specification" (identical to ISO 23950) is of interest since it was recently adopted for all federal and national information in the U.S. (within the framework of GILS, Government Information Locator System). Possibly authorities in Sweden and many other countries will follow the path, thereby offering the user the possibility to search and retrieve information from a wide range of Z39.50 compliant databases and other repositories. (See Mikko Kalliosalo, *Framtidens verktyg för informationssökning? En studie av informationssökningsprotokollet Z39.50 för Kungliga Biblioteket*. KTH, Nada, TRITA-NA-E9749, 1997. For information on the standard see the Library of Congress Maintenance Agency page at http://lcweb.loc.gov/z3950/agency/.) A combination of Z39.50 and XML is a most promising direction for future development of information retrieval at the Internet.

A fifth direction is to remedy the annoyingly immature technology for implementation of "consumer guidance" on content, especially quality control and relevance control. A content designer should aim at incorporating such meta-data into the content itself. A teacher who downloads material from a courseware archive on the Internet or on a CD-ROM needs information on the quality of the material -- has it passed refereeing? how accurate is the proof-reading, measured by number of anticipated faults per 1,000 characters? --, on its provenance, on the revision history, on the intended school level, etc. Further, he or she would be helped if this meta-data formed part of the downloaded material itself instead of being tied to TOCs, indexes and search mechanisms available only on-line at the web site or only as long as the CD-ROM is spinning. The international project TEI (Text Encoding Initiative) has developed a scheme, called the "TEI header", to furnishing information objects with this kind of portable meta-data. Many other attempts to standardise meta-data are going on in the archive, library and museum communities and elsewhere. One important aspect is that digital archives should be designed in such a way that meta-data might be inspected and managed separately. The user should not be forced to load the rest of the content. If the archive is rich and complex this calls for sophisticated graphical presentation tools that enable the user to navigate through and manage the meta-data in order to identify, download and reorganise an appropriate sub-set of information.

Content design in the sense described already takes place in corporations and organisations where for example editors or technical writers pool information content in portable format to be shared by collaborators and other users within the organisation. For anyone who hopes for the promises of integrated open hypermedia to be fulfilled, it is sad to witness that at a later stage this information more often than not is delivered in crippled format, intended for a specific delivery platform. Thus users can't benefit from all the value added in the in-house production. Take for example the available electronic versions of the Swedish law texts as published by Norstedts and by Fakta Information. Both publishing houses use SGML markup in the editing and pre-press stage and for their in-house management and updating, while they ship CD-ROMs where the information is fettered to a proprietary presentation format. It is easy to understand the commercial logic in this phenomena. After having invested something between 25,000,000 and 30,000,000 SEK in preparing the content for the Swedish Karnov edition, it is understandable that Fakta Information hesitate to make it available in portable format. Still, there should be other possibilities for publishing companies to earn money than to destroy information. Perhaps some fundamental cultural, scientific, or legal resources could be made freely available in digital portable format, like the air we breath, thereby enabling the information industry to make profit out of added value: updates, commentaries and extensions, refereeing services and quality stamps, educational services etc. To keep to legal publishing as an example, an ongoing project within the European Union aims at publishing all member states' legislation on the Internet. In Sweden the DORIS group has for a long time had the same ambition. An example from another field is Skoldatanätet, established by Skolverket (the Swedish National Agency of Education). Skoldatanätet procures shared resources available to all Swedish teachers and pupils. It is at present exported to other European countries. Such initiatives could, if developed in the direction of integrated open hypermedia solutions, lead to new business possibilities for the information content industry. If a sufficient critical mass of high quality content in portable formats did exist in a certain field, for example in legal information or education, there should also be a marked for add-on products from commercial content providers.

4. POSSIBLE INTERESTED PARTNERS

The proposed CID project might possibly be of interest to information providers as publishing companies or organisations as Skolverket and Handikappinstitutet. Some technological aspects are of relevance to many branches of industry, like the current development of markup meta-languages (especially XML) for content distributed on Internet or intranets. CID's labour union partners might pay attention to the fact that the Danish LO (the blue collar labour union) has developed an extensive system for content delivery to their members which was said to have been the worlds largest SGML-project so far.

Finally, the proposed project might inform and be informed by other ongoing and planned CD-projects which are more focused upon presentation technologies, user interface design and visual arts.