ECHO
EUROPEAN CULTURAL HERITAGE ONLINE

Web of Culture and Science

An Open Source Project Proposal for Information Management in the Humanities

Jürgen Renn and Peter Damerow (MPIWG)

in cooperation with
Malcolm Hyman (Harvard University),
Mark Schiefsky (Harvard University), and
Urs Schoepflin (MPIWG)

The Web of the future

The Web represents a powerful achievement in the connectivity of human knowledge that until recently seemed unconceivable. The revolution it has caused has rightly been compared with those of the invention of writing and of the invention of printing technology. But the rapid development of the Web itself is about to surpass the basic development of the technologies it is based on. In its present form, the Web makes more promises than it can actually keep – at least as long as it is restricted to the specific paradigm that has originally given rise to it. In particular, it lacks longivity, interactivity, and transparency.

Just as it was the case when the Internet was created by turning a network of computers into a medium representing a universal hypertext, also its future as we see it will rather depend on requirements and possibilities that are revealed only in the context of innovative usage scenarios. Such usage scenarios will emerge when the Internet is used as a virtual public think tank, a web of culture serving as a medium of reflection on current global challenges of human civilization such as the destruction of ecological equilibria, social impacts of epidemics and drug addictions, or terrorism and other devastating consequences of oppression and increasing mass impoverishment.

If both the natural sciences and the humanities should even under these conditions of global challenges be capable of providing the knowledge crucial for solving the problems of the human species, then this knowledge must also be represented, integrated, and made available in a form allowing for global orientation and action. It is time to take up the opportunity offered by the Internet to create such a medium of global human reflection.

Berlin, January 2003

As yet, nobody knows whether or not the net will be developed in this direction and what precisely it will look like if it does become such a public think tank, but some requirements are evident. A future web of culture

- will have to preserve and offer free access to the cultural heritage of mankind against ruthless pragmatism of technical and economic progress;
- will have to provide public facilities and equal opportunities for computer-assisted training and education;
- will have to encourage the free exchange of information and arguments across still existing social, political, and religious boundaries;
- will have to garantuee an enduring memory of mankind comprising a representation of human history in a new form;
- will have to develop mechanisms for self-organisation and for the evaluation of the information it makes available.

For realizing such a web of culture it is not sufficient to make the current web just more efficient and to extend its areas of application. Such a realization requires moreover content-driven innovations as they will emerge as soon as the challenges described above are seriously taken up.

Overcoming the current limitations of the web-processing of texts

Due to its origin in the idea of hypertext, the World Wide Web is centered on textual data enriched by illustrative insertions of audio-visual materials. The status quo paradigm of the Web is a client-server interaction, that is, a fundamentally asymmetric relationship between providers inserting content into the Web hypertext (server) and users who essentially read texts or provide answers to questions by filling out forms (clients). The hyperlinks of the Web represent structures of meaning that transcend the meaning represented by individual texts, but, at present, these webized structures of meaning, lacking any longivity, can only be blindly used e.g. by search engines which at best optimize navigation by taking into account the statistical behavior of web users. However, these meaning structures can so far hardly be made themselves the object of interventions by the web community. There is at present no way to construct complex networks of meaningful relations between web contents. In fact, the providers have no influence on the links to the contents provided by them and the users have no impact on the available access structures to the content, except by becoming content providers themselves.

This asymmetric client-server relation largely determines the functionalities of the existing web-software. Webservers are not the standard tools of users, while the web-browsers used by them are restricted to accessing existing information in a standard form with only limited possibilities of further processing that information (such as e.g. changing fonts and background colors). As a consequence, the present Web offers no possibility for (radically) different views of the same underlying content, depriving users from the creative potential inherent in the dynamics of the ever-changing Web hypertext.

The Web of the future will thus continue to be essentially based on the representation of meaning by text. However, contrary to the existing web, its emerging paradigm is no longer constituted by the client-server assymmetry but by informed peer-to-peer interactions, that is, by a cooperation of equally competent partners who jointly act as providers and servers at the same time. Future users will work on shared knowledge by constructing new meaning while accessing the existing body of knowledge represented in the Web through meaningful links to texts and text corpora. An important framework for creating such meaningful links can be provided by what is presently discussed as the „semantic web,“ that is, the automated creation of links between machine-understandable metadata, using RDF technology. In a further perspective, however, such semantic linking will not be restricted to the use of specifically prepared metadata sets but will exploit the meaning structure of the Web itself in order to provide a content-based semantic access to information.

The proposed project

The proposed project will be pursued in the framework of ECHO (European Cultural Heritage Online), an initiative aiming at brigding the gap between social sciences and humanities and the new information technologies in order to establish a new quality of access to cultural heritage on the Web, thus transforming the latter from an ephemeral communication network of providers to an enduring representation of the shared knowledge embodied in the cultural heritage of mankind.

However, in order to represent culture adequately on the web, not only a reorientation from the currently prevailing use of the net for commercial and topical issues to its use as a publically accessible comprehensive knowledge system is required, but also the development of new technologies supporting such a reorientation. The functionality of the software to be developed will accordingly depend not so much on exploiting technological potentials immediately at hand such as the increase of mass storage capacity or of transmission rates but rather on the requirements of the content to be represented and processed. It is the general aim of the proposed project to develop an adequate Internet technology while applying it, at the same time, to the content of ECHO serving as a testbed for this development. More specifically, the software to be developed will constitute a test platform for technologies to be included in “browsers” of tomorrow which might then more adequately be designated as “knowledge weavers.”

The design of this software platform which will be specified in the following is widely based on experiences and technical developments already accumulated by the leading institutions of the consortium responsible for the webprocessing of textual data. In particular, in the context of the Archimedes Project, jointly pursued by one of the institutions of the consortium with three US American institutions, new tools for distributed content-based work with XML documents have been developed and applied to large corpora of historical documents written in various languages. Examples are the downloadable content-based XML browser and annotation environment Arboreal, the web-based morphological analyser Donatus, and the tool for accessing distributed dictionaries Pollux (http://archimedes.fas.harvard.edu/). The developed technology has sofar been mainly applied to documents in the history of mechanics (http://archimedes.mpiwg-berlin.mpg.de/) and to cuneiform accounting documents (http://cdli.ucla.edu/progress/arboreal.html). These encouraging achievements will constitute the point of departure for creating the planned software platform.

This planned software platform will be characterized by displaying functionalities and addressing issues as they are briefly surveyed in the following.

- *Basic data structures.* It is assumed that the current html-format centered on the display of data is already obsolete and will, in any case, be replaced by XML data formats more adequate to their content structures. Accordingly, the software platform to be developed will consequently be based on the processing of XML documents exploiting the hitherto unused potential of this standard of the future. The XML format allows for the qualification of information by tagging it in the framework of content-specific document type descriptions. Such tagging is the prerequisite for content-based information processing of textual data. The only reason that this potential of the XML format is currently still rarely used within Internet technology is the lack of adequate software instruments as they will be developed in the proposed project.
- *Data transformations.* Whereas current browser technology does not allow for any other use of html tagging than changing the mode of text display (font, color, etc.), the new platform to be developed has to provide facilities for processing the browsed data according to the content they represent (e.g. to create word lists, extract specific information, attach the content to terminologies such as the entries of encyclopedias etc.). Such data transformations and linking mechanisms will be realized in a modular way with plugins using XSLT technology.
- *Natural language technology.* The XML format allows, in particular, to process textual data in dependence of their language. The software platform to be developed will use this potential by systematically applying language technology and by enhancing its range and scope. In particular, it will allow the integration of already existing distributed language resources within local data processing such as morphological analysis and linking to dictionary entries. As far as they are not yet available, techniques of morphological analysis and electronically accessible dictionaries will be provided for all European Union languages. Software modules representing this technology will be realized not only as local plugins but also as web plugins enriching the Web as a whole. A formal language-specification language will be developed supporting the continuous integration of further natural languages. Furthermore, a software architecture will be established which allows for the integration of richer models of natural language structures. Such models will make it possible to develop methods of context-sensitive morphological analysis for the morphological disambiguation in syntactic contexts, to deal adequately with discontinuous constituents such as compound verbs, or to automatically resolve pronominal coreferences. The global implementation and improvement of language technology will thus provide the necessary precondition for identifying, accessing, extracting and processing specific contents of text documents independent of their representation by particular languages. This content-driven technical innovation will thus foster the development of a fundamentally new way of semantic linking in the Web.
- *Formal language technology.* Many scientific disciplines, in particular those of the natural sciences, make extensive use of formal languages such as those embodied in mathematical and chemical formulae. Powerful programs for symbolic processing of the information encoded in these formulae are available. But they are, in fact, not as widely used as they could be if the full potential for a web-based distributed storage and processing of such information would actually be realized. In spite of the connectivity of the existing Web, the representation of such information in web-inadequate formats and the prevalence of proprietory software solutions make it in fact impossible to connect the distributed resources.

  The future perspective of formal language technology, on the other hand, will be shaped by recent developments to successfully apply XML formats to encode not only natural languages but also formal languages. The introduction of the MathML and ChemML specifications have opened up the possibility for dealing with mathematical or chemical formulae in a way analogous to information encoded in natural language. The planned software platform will accordingly provide the means for handling and displaying such new formats and make it possible to identify, access, extract, and process content encoded in formal languages.

Berlin, January 2003

- *Data on data.* A basic faculty of human thinking is the ability to reflect on existing knowledge and to produce, so to speak, data on data. The outcome can be highly complex as in the case of theories reflecting and overcoming the limitations of other theories, thus constituting a network of meaningful relations. Although the representation of such networks of reflection in the Web does not raise unsolvable technical problems, the very mechanism of such a network of reflection can only to a very limited extent be mapped onto the structures of the Web due to the incompatibility of its traditional paradigm with the creation of meaningful relations between contents in an adequate way. Even when considering the future of the Web, the problem of the reflection on Web content is predominantly perceived as that of creating, adequately representing, and exploiting metadata in a limited sense. As a particular consequence, the generation of the specific kind of metadata produced by scientific disciplines through analyzing, annotating, and reformulating the content of texts is not supported by the present infrastructure of the Web and not even adequately represented by corresponding relations between textual resources in the Web. The software platform to be developed will support the creation of such metadata by interactive working environments for scientific analysis and annotation, resulting in turn not only in qualitatively improved relations between text sources but also in richer metadata sets for navigation through scientific contents and their mutual relationships to each other. Building blocks of such working environments are, for instance, technologies for clustering parts of texts and tools for extracting terminologies and using them for content driven navigation through text sources. Applying the presently available techniques for handling metadata as well as those currently under development to the enriched, content-specific metadata structures thus produced, will provide a new infrastructure for web-based scientific research and collaboration.
- *Content-sensitive linking*. The new infrastructure of the Web which results from a realization of the explanatory power of data on data provides a solution for a serious problem of the current as well as of the future Web, that is to produce order in the ever-growing complexity of the Web by content-sensitive linking. Even the most tricky search engines will reach their limits as long as the search criteria can at best exploit the statistics of human decisions about the quality of data. If, however, navigation can be based on content-specific metadata resulting in dynamically changing ontologies, and if powerful link editing functionalities become part of the future "knowledge weaving web environments," a self-organizing mechanism of the Web will be implemented which in a middle-range perspective will improve the hypertext linking of the Web. The software platform to be developed will support this process by integrating RDF technology for representing and editing of ontologies, continually modified by the Web users. These linking tools will include not only relations between textual data but also instruments for creating and editing links to knowledge represented in multimedia (text, image, video, audio).
- *Culture in the Web.* The cultural heritage put online by the ECHO project will serve as the testbed of the software platform to be developed. A number of project activities will be devoted to the realisation of this testbed function. In particular, a typology of the multimedia cultural heritage content will have to be developed in order to specify the requirements that will have to be met by future developments aiming at the improvement of the Web as a whole. Document specifications will be written for the major types of cultural heritage data covered by ECHO. These document specifications will be described in a formally defined, computer-readable description language to be developed. This "Document Description Language" (DDL) will make it possible to specify not only the ontology of a document, but also its formal structure, as well as the characteristic interface to human activity (presentation and editing). For writing document specifications in the DDL, user-friendly frontends will be developed in the form of palette-based graphical user-interface tools. Based on the document type descriptions, a modular display environment for the cultural content of the ECHO project will be developed. In order to develop into the backbone of the cultural memory of Europe, the content-based, dynamically evolving typology has to become the core of a sustainable archive solution of data and its transparent data architecture. The RDF technology applied to content-specific metadata will secure that this archiving will not resilt in a frozen data repository but give rise to an infrastructure for a living history of culture.

Berlin, January 2003