MPIWG proposal within the MPDL framework

The Max Planck Institute for the History of Science (MPIWG) proposes a project within the Max Planck Digital Library (MPDL) framework. With the MPDL as a central body for scholarly information management, the Max Planck Society (MPG) has established a unique structure to take up and further generalize tools and services that have been developed at individual Max Planck Institutes and to make them available for the benefit of the entire MPG. This relationship—research-driven development at the research front on the one hand and generalization by the expertise of a central body on the other—is a mechanism ensuring that the advanced services made available at the MPG have proven their specific relevance for research and can thus be expected to have an immediate impact on the work of the scholarly community.

Our project has been carefully designed with this objective in mind, i.e. to allow for being taken up and incorporated into the set of tools currently under development at the MPDL. Although it responds to urgent research needs at our Institute, it by no means represents an isolated development. We rather foresee that it will meet clearly articulated requirements in humanistic research and in other scholarly communities that deal with textual sources as well as image and geospatial data.

Project Description

This project aims at developing four complementary services within the MPDL framework. The services will be *prototypes,* which can be further generalized within the MPDL, if there is sufficient demand for them.

The four services will constitute (1) a workflow for developing texts in an XML format that represent historical (printed or manuscript) sources; (2) a content-based access mechanism for these texts that incorporates language technology, which will be built upon the MPDL infrastructure and will be publicly web-accessible; (3) software for Virtual Exhibitions; and (4) an Open GI (Geographic Information) network for the retrieval of scientifically relevant geo-information.

Rationale: The digital humanities need large-scale content acquisition/digitization, a complete pipeline for the transformation of raw data to structured XML, the establishment of editorial (meta-) conventions, and a thorough review of content structure in the context of use. To allow for the free flow of communication on which global scholarship depends, such texts must be open access and must be presented in an environment that maximizes their scholarly utility. Such an environment must offer *content-based* access to the texts, which includes sophisticated search capabilities that depend in part on natural language processing (NLP). Virtual Exhibition software allows for the publication of interactive online content (including text, images, video, and user-interaction elements) for the general public. Archaeological and historical map data have only in very few cases been digitized, owing to the complexity and proprietary nature of most GIS software. An Open GI network will provide a high quality, standardized, and open access means for the publication of geospatial data of scientific relevance that can be used directly by scholars, without extensive technical support.

The project we propose differs from large-scale commercial endeavors (e.g. Google Books) and is designed to be complementary to such endeavors rather than a replacement for them. Such endeavors have as their primary aim the digitization of vast collections of printed materials, with the aim of making the materials available to the public. Only rudimentary search facilities are provided; these are based on full-text indexing of unstructured text derived via OCR. By contrast, our aim is to make a carefully selected set of resources available to scholars, with sophisticated content-based access and support for interactive distributed research. Similarly, the Open GI network focuses on

an application that will allow researchers to share geospatial research data freely and to integrate these data with other information sources. Although the Open GI network makes use of Google Earth, no critical functionality will depend on any commercial provider, and the system will be capable of inter-operating with open systems such as NASA World Wind.

ECHO Project

The ECHO (European Cultural Heritage Online) project,[1] coordinated by the MPIWG, was funded in 2002 as a research, technological development, and demonstration (RTD) activity within the Fifth Framework Program of the EU. Today ECHO contains material provided by about 50 institutions around the world, including material related to: anthropology, archaeology, botany, demography, geography, history of art and architecture, history of science, language study, literature, philosophy, psychology, and religious studies. Current holdings include ca. 230,000 high resolution images of documents and artifacts, ca. 57,000 page transcriptions in XML, and ca. 240 video sequences.

The current technical infrastructure of ECHO is inadequate for indefinitely maintaining a (growing) collection of this size. Within the MPDL framework we intend to pilot a replacement architecture for ECHO as well as to prepare a migration path for the ECHO content.

ECHO is closely linked with the services discussed elsewhere in this proposal. Within ECHO, the basic object is an XML transcription of printed or manuscript materials, combined with digital page images of the original. Language technology and automatically provided links to reference sources enrich the digital presentation.

Existing tools and tools under development

Tools already developed at the MPIWG that are relevant to the services enumerated above include:

- **Digilib**, a web-based image presentation tool that can retrieve part or the whole of an image from a data store that includes multiple resolution versions of an image (some generated dynamically). With Digilib, the user can flexibly navigate, zoom, scale, and apply transformations to the image within the browser. Relevant portions of the image data are sent over the network at the appropriate resolution. In addition, Digilib allows the annotation of areas of interest within an image and the sharing of such annotation sets.
- **Donatus** is middleware designed to present through web services a common interface to NLP tools. Work so far has focused on word segmentation, orthographic normalization, and morphological analysis. Future work will include the incorporation of higher-level tools, e.g. (shallow) parsers.
- **Pollux** is a distributed online dictionary server, designed to publish any sort of material in which an entry is commonly retrieved by a *headword* or *keyword.* Dictionaries may be locally hosted by a Pollux server, or Pollux may serve as a proxy that sends a request to an external dictionary resource (e.g., Das Digitale Wörterbuch der deutschen Sprache des 20. Jh., Electronic Pennsylvania Sumerian Dictionary).
- **Arboreal** is a new type of "browser," based on XML. It supports flexible tree-based navigation of XML documents, NLP technology via Donatus, powerful search functionality (including XPath, searching over an arbitrary subtree of the document, powerful regular expression search, orthographically normalized search, lemmatized search…). Arboreal offers moreover the ability to work with multiple texts in parallel (e.g. texts plus commentaries and translations; multiple editions); a parallel text (such as a translation,

1 http://echo.mpiwg-berlin.mpg.de

commentary, or notes) can also be composed directly within Arboreal. Arboreal in addition provides rich annotation (based on out-of-band XML), with especial strength on the annotation of *technical terminology*.

- The **Virtual Exhibition** is open source software to present exhibition and museum content publicly on the Web. In the virtual exhibition model, roles are maximally independent. Content creators do not need to be concerned with design or technical implementation issues, and graphic designers do not need detailed knowledge of either the underlying software or the scholarly/scientific context. Work proceeds in a networked environment where content creators can easily use a web browser to create "slides" with textual content as well as audio and video content that is hosted in a shared web-accessible database. The Virtual Exhibition software was originally developed for the exhibition "Albert Einstein: Chief Engineer of the Universe" (2005). Since then, it has been used in other projects at the MPIWG as well as by other institutions, including the University of Pavia, the Fundación Canaria Orotava de Historia de la Ciencia, the Kroppedal Museum for Astronomy, and the British Museum.
- Prototype work on the **Open GI network** has been performed at the MPIWG by Sebastian Schröder as part of a *Diplomarbeit* "Web 2.0 und der Einsatz in der Wissenschaft" (in press, 2007). These prototypes make use of Google Earth and KML to provide geospatial data access for the Cuneiform Digital Library Initiative (MPIWG/UCLA) and for joint work of the Independent Research Group led by Dagmar Schäfer and the Chinese Historical Geographical Information System (CHGIS) hosted at Harvard University.

Further information about these tools is available at the following URLs:

- http://developer.berlios.de/projects/digilib
- http://archimedes.fas.harvard.edu/
- http://www.einsteinausstellung.de/
- http://cdliwiki.mpiwg-berlin.mpg.de/doku.php/web_2.0_und_der_einsatz_in_der_wissenschaft

Planned work

XML Workflow Service

This service focuses on the preparation of structured XML "editions" corresponding to printed or manuscript resources from semi-structured data entry (typically procured from commercial services). Critical tasks include: correction of errors in the transcription, markup of the structurally meaningful divisions of the work, markup of basic "semantic units" (sense units approximated by the sentences of a modern printed edition), and correlation of the indicated page boundaries in the transcription with digital page images. Tools for some of these tasks will be offline, but we believe that online tools would constitute an attractive alternative in a number of cases. During this workflow process, the MPDL storage layer will serve as repository for the document. Tools for facilitating the workflow tasks will be developed within this project. These tools will draw upon preexisting technology developed within the MPIWG; e.g. the morphological analyses provided by Donatus can be used to flag probable data entry errors.

It is our expectation that the workflow developed here will be useful to scholars working in many humanistic, as well as some scientific, fields.

Content-based Web Access

The primary method of disseminating texts will be to present formatted pages of the XML transcription in parallel with digital page images. (Digilib will play the primary role in the disseminating the latter.) Here a basic subservice will extract a given page from an XML fulltext and provide a balanced (or "symmetrized") version. Such a subservice is needed, since the XML between two page break milestones usually is not a well-formed XML fragment without further processing. The system is designed to support multiple XML vocabularies—which will require minimal configuration information—such as the TEI document type or ECHO document type. Subsequent to extraction and production of a balanced XML fragment, the display pipeline involves the following three major steps:

- **Rendering.** Rendering of the balanced XML fragment will be performed with XSLT on the server side, yielding XHTML for the client. XSLT will be readily pluggable, allowing for multiple output options.
- **Enrichment.** The generated XHTML will be enriched with: inline images, links to external resources (e.g. Pollux dictionaries via lemmatization provided by Donatus; geospatial data). At this stage, transliteration of various sorts is also possible (should a Greek text be displayed in a Romanization or in Greek characters? should an Arabic text be displayed fully voweled, in its typical rendition, or in Romanization? should a Sanskrit text be displayed in Devanagari, or Tamil, or Romanization, or IPA? a Chinese text in traditional characters, simplified characters, or pinyin?). This is also the layer at which named entity resolution is most appropriately realized.
- **Generation of a synthetic view.** The XHTML view will be synchronized and presented in coordination with the appropriate digital image, provided by Digilib.

In addition to the basic display environment, a language-sensitive indexing tool needs to be constructed. Such a tool will allow searching a particular text, a corpus, an arbitrarily selected group of texts/corpora, or all texts for one or more natural language words. The search functionality will be developed using an open-source tool (e.g. Lucene) in combination with the NLP technology hosted by Donatus. Thus, for instance, it will be possible to search for all inflected forms of a Latin verb (or only a subset of those forms).

There will also be support for accessing texts through human-constructed indices, which reference the texts through XPointer. In this way, scholars will be able to develop an access approach to a given text.

A further component of this project is to extend the Arboreal browser to be able to make use (both read/write) of the MPDL repository. This extension will provide scholars with an alternative/complementary access modality. In addition, Arboreal, which is an inherently network-neutral application, will be able to offer storage within the MPDL repository as an alternative strategy for saving content generated within the program.

It is also our intention to integrate a general statistical toolkit currently under development by the Scholarly Computing Group of the MPIWG into this framework.

The Virtual Exhibition

The Virtual Exhibition environment is intended to allow for the easy creation of web-based exhibitions by museums and other institutions interested in the public presentation of scientific, historical, or cultural heritage materials. We plan to extend the tools that allow "virtual spaces" to be defined—such as the floor-plan of a museum, the topography of an island or a garden, or the layout of an archaeological site. Such spaces are navigated by visitors in exploring the textual and

multimedia content of a virtual exhibition. A more flexible approach to the definition of virtual spaces and creation of hyperlinks within these spaces will allow content providers greater creative possibilities and will make the software attractive to further institutions. We intend in addition to re-architect the system to make it more modular and more easily distributable and installable. An additional area of work will be increasing integration with the other services described in this proposal.

Open GI network

The Open GI network is based upon the open source PostgreSQL database with the PostGIS module. Satellite and aerial photographic data will be provided by a flexible choice of commercial and open access vendors. The Open GI network will make use of public API functions of the MPDL, and will in turn provide a public web services interface. The goal is to make GIS work manageable for individual researchers, for whom the Open GI network will constitute a valuable tool for research, publication, and presentations. The Open GI network will be integrated with other digital library projects developed at the MPIWG.

Personnel

The MPIWG has for many years invested, and continues to invest, significant resources of its own in the development of scholarly computing and digital library applications. It has also invested significant resources and time in specifying the usage scenarios that underly current development activities on the Scholarly Workbench component of the eSciDoc2 framework. The purpose of this proposal is to transform this investment into concrete products that will be of utility to the scholarly community.

For the activities described in this proposal, the MPIWG asks for 3.5 full-time positions for a period of 2 years. Two programmers (one with expertise in geospatial systems and graphics, and one with expertise in text and language processing) are needed, as well as a content coordinator and an assistant (½ position).Programmers will work on software development (in consultation with the coordinator at the MPIWG, Malcolm Hyman) and with feedback from personnel in the Scholarly Computing Group of the MPIWG. The main programming tasks include the development of tools for XML text structuring, correction, and enrichment; development of modules of the display environment, including the interfaces with Digilib, Donatus, and Pollux; incremental improvements to the Donatus and Pollux services; enhancement of the Virtual Exhibition software; and development of the Open GI network. The content coordinator will work on content integration and scholarly coordination related to the migration of the ECHO project. The assistant will work with the project coordinator in workflow development and internal/external communication activities and will be responsible for testing and quality assurance. Personnel will cooperate with the MPIWG Library and with scholars at the MPIWG in developing the technical and social aspects of workflows and with personnel at the MPDL and FIZ in integration and compatibility assurance.

Conclusion

The MPIWG is enthusiastic about the potential of further developing its innovative research tools in the context of the MPDL infrastructure. We also anticipate that the services and tools developed within this project will be of considerable interest to a broad community of scholars.