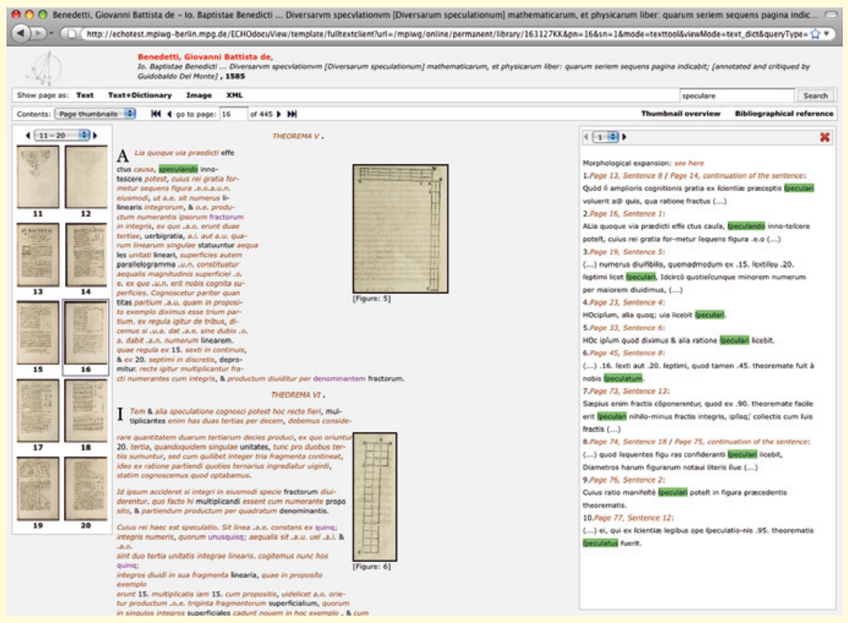
Department I

The MPDL-MPIWG Project

XML workflow from a physical book to a fully marked-up text viewable in a browser

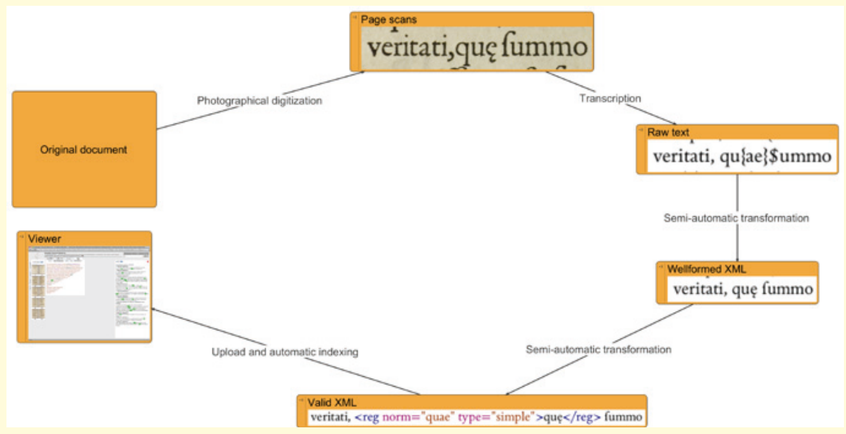
This project aims at the creation of a workflow to digitize historical sources (Western as well as Chinese) and to present them in a state-of-the-art display system with connection to other resources such as dictionaries or Geographic Information Systems (GIS).



The display system shows a text from Benedetti (1585) with both a text and a dictionary view (bold words are clickable). To the left, thumbnails, the list of figures or the table of contents can be toggled. To the right, a search window can be popped up.

DESpecs: from digital images of a physical book to raw transcript

The DESpecs (Data Entry Specifications) comprise a set of rules on how a book from the Early Modern period should be transformed using a minimal markup language and the whole set of Unicode characters. In this way, structural elements like chapters, the position of images or even tables can be easily transformed into a simple text file which can then be processed. The DESpecs can be enhanced to accommodate books from other periods. In addition, special instructions can be formulated for the peculiarities of individual books.



Schematic overview of the workflow. Digital scans are provided, for example,by the Institute's library.

Transformation to XML and validation with Echo Schema

After having checked the transcript for errors, the raw text version is enriched with additional information such as metadata; links to the respective page images are added at each page break.

The following transformation to an XML document is divided into two steps. In the first step, the raw text is converted to a well-formed XML document. For this conversion, tools have been developed that support the scholarly work. Manual work is only necessary for semantic markup.

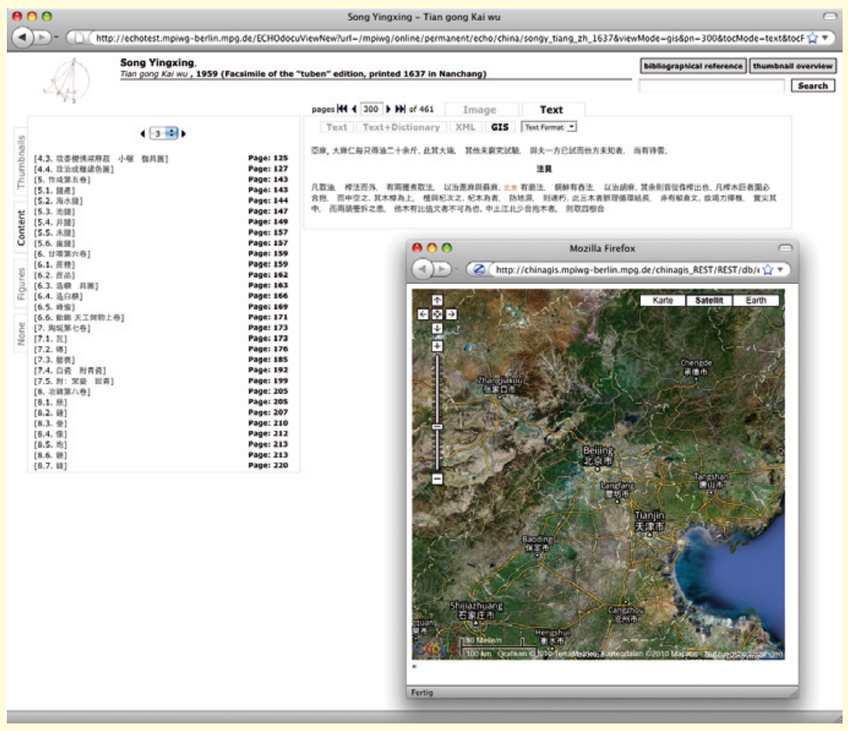
The second step validates the document with a set of XML markup rules, which were also developed by the project group. The purpose of these rules is to ensure that all texts share the same set of tags and properties. This is important for the integrity of the data and for transformation into the format displayed in the browser.

Display system

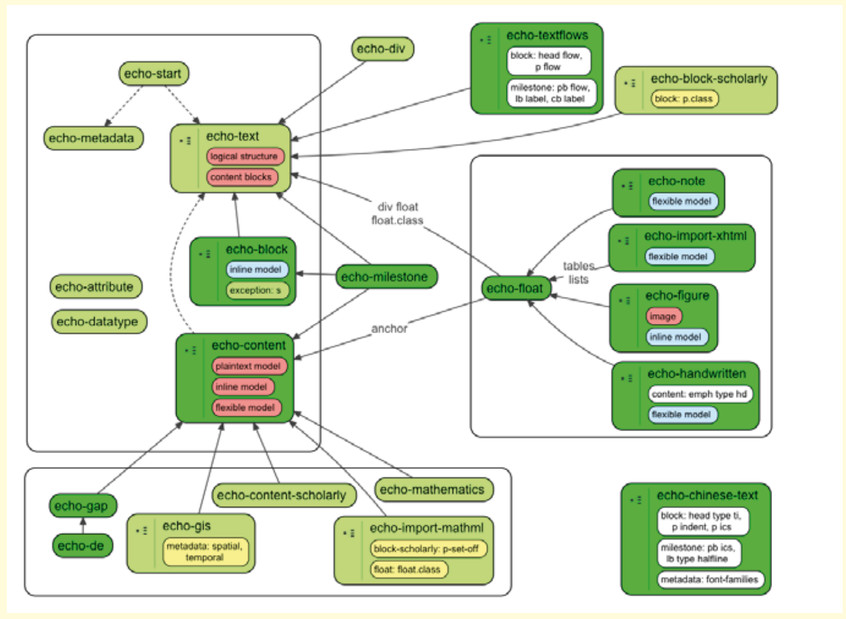
Based on the markup in the XML document, the display system allows the transcribed text to be shown in synchronization with the page images. Figures in the text are shown as well, and the text can be downloaded as XML or PDF. During the upload of the document into the display system, the XML text is analyzed using language technology. Thus, the search for a word will not only find this word, but all other forms of this word as well. This mechanism is also used to generate links to online dictionaries.

141.**Page 25, Sentence 9**:
Pa-tet enim ex præcedenti theoremate .m. q. radicem eſſe quadratam producti .l.e. in .e.p. quod productũ ſit (...) .c.g. cogitemus pariter duo quadrata .l.e. et .e.p. eſſe pariter corpo-rea, tantę profunditatis, quantam, vnitas linearis (...)
142.**Page 26, Sentence 2**:
Vn-de ex decimaoctaua, aut decimanona ſeptimi, eadem erit proportio .a.g. ad .c.g.quæ eſt .l.b. ad .l.x. corporeum, ſed ex .25. vndecimi & prima ſexti, ita (...)
143.**Page 26, Sentence 4**:
Vnde ex nona quinti .a.K. æqualis erit .e.b. & conſequenter æqualis .m.e.
144.**Page 26, Sentence 5**:
Iam verò ſit .u.g. productum .l.b.cubi, in cubum .o.p. vt ſupra dictum eſt (...) duarum propoſitio-num, decimaoctaua, aut decimanona ſeptimi, eandem futuram proportionem .u.g.ad .a.g. quæ eſt .o.p. ad .x.p. quadratum corporeum.
145.**Page 26, Sentence 6**:
Quare ex poſtremis, dictis ratio-nibus, eadem erit proportio .u.K. ad .a.K. quæ eſt (...) q. ſicut nume rus .q.e. ad ſuam vnitaté, ſed cũ numerus .a.K. æqualis ſit numero .m.e. vt probatũ eſt erit ergo ex vndecima & nona quinti, numerus .u.K. æqualis numero (...)
146.**Page 26, Sentence 7**:
At .f.g.pariter æqualis eſt numero .m.q. ex præcedenti theoremate, vnde .K.u. pariter æqua lis erit .f.g.
147.**Page 26, Sentence 8**:
Itaque ſequitur .u.g. cubum eſſe, & f.g. radicem ipſius, æqualem numero .m.q. quod quærebatur.

Detailed view of a search window. The search term was "esse". Instances written with a long 's' are also found.



The display system shows a Chinese text which has already been marked up with place names. If "GIS" is selected in the display system, the place names are highlighted. A click on it opens up an interactive map showing the location.



The structure of the XML validation schema.

contact:
Jochen Büttner
buettner@mpiwg-berlin.mpg.de

project members:
Andrey Bukhman
Robert Casties
Malcolm D. Hyman

Wolfgang Schmidle
Klaus Thoden
Josef Willenborg
Dirk Wintergrün

MAX PLANCK INSTITUTE
FOR THE HISTORY OF SCIENCE