From Text to Topics — Zigzagging Towards the Knowledgebase of Tang Civilization

***Christian Wittern** (wittern@zinbun.kyoto-u.ac.jp)*
Kyoto University

Abstract

A few years ago, the prospect of having access to a large amount of digitized data promised to give a completely new direction to the field of Chinese Studies. Although today we have such databases as the Siku Quanshu (四庫全書,) Fulltext Database[1], as well as many other texts, some of them even freely available on the Internet, the benefits of this has been limited. There are many reasons for this, not all of them technical. Of the technical reasons, the limited, idiosyncratic interface that each of these database provides, and the unstructured data it operates on are probably the most important ones.

The *Knowledgebase of Tang Civilization* is an attempt to remedy this situation, at least for material relating to the Tang, by providing a comprehensive electronic archive of information about China during the period of the Tang dynasty (618-907 A.D.) in a way that allows new ways to access, analyze and expand the information. Work on this Knowledgebase started in 2003[2] with initial funding for 5 years. This presentation will present some of the experiences gained in the first development phase.

The design of the Knowledgebase uses a two layer model, that distinguishes the 'information layer' from the 'resource layer'. The organization of the information layer is based on the topic map paradigm.[3] to allow for the expression of ontology subtrees, with links from the information layer back to the resource layer, which will hold primary sources.

Its main point of access for researchers will be a web application, but other interfaces will be developed.

Initially most of the information to be included will be textual, but will in due time be enhanced by images, visual reproductions of objects, digital maps and animations of events. The distinguishing feature of the knowledgebase is the way information items are interconnected in a flexible and innovative way.

The information in the knowledgebase will be organized along the following information axis:

- Personal names, dates and activities of people of the Tang.
- Placenames and georeferences to there locations, administrative geographical units, digital maps.
- Works created during the Tang, including texts, artefacts and buildings
- Calendar and time
- Events of importance and influence

Obviously, many if not all information items will be accessible through more than one of these axes; internally they are cross-linked and form more of a web-like structure. Additionally, these items are organized in hierarchical ontologies. This allows to access the information also based on their position within the hierarchy, or on the relation with other items. For geographical locations, like a city for example, such a hierarchy would consist of the upper administrative units it belongs to; for persons this could consist of the family line, but also the region of origin, the school or tradition of thought, in the case of monks also the ordination line and line of transmission.

The challenge in the first phase of the development, which will be concluded by the time this presentation will be given, was to design a way to bootstrap the Knowledgebase. For this purpose, two dynastic histories (the *Jiu Tang Shu* (945) and the *Xin Tang Shu* (1060)) and one chronologically arranged historical account by Sima Guang (*Zizhi Tongjian*, 1084) have been chosen to provide a basic set of information about the Tang period. This idea relies on the fact, that the dynastic histories do not only provide a day to day chronicle of court affairs and other events, but also include monographs on a variety of subject matters, including geography (with detailed accounts of administrative units, their changes in size and denomination, local production, population etc.), calendar (including accounts of the calendar systems in use), ritual observances, music, astronomy, offices, state finances, law and a detailed bibliography of works known to have written in that period. In addition to that, more than half of the text of the official histories is taken up by biographic accounts. In the case of the Tang, there are two such histories, since in the eyes of Ouyang Xiu, the editor of the second, "new" history, the first one had some defects in style and presentation.

The texts are encoded in XML using the TEI vocabulary. In a first phase, only structural encoding was applied, so that the texts could be accessed using XML technologies[4] and could be further processed. It was then started to add semantic markup to allow for automatic extraction of information.

It should be obvious, that this collection provides rich material that could be mined for inclusion in the Knowledgebase, but the challenge was to find an efficient way to mine that information, generate topics from them and relate them to each other in the way outlined above. The presentation will focus on the strategies employed and results achieved and will then try to look at how to generalize these methods. It is also planned to show a prototype of an interface, that allows further enhancement of the data.

1. A database published by Chinese University Press, that includes on 176 CD-ROMS an electronic text of the anthology Siku Quanshu, which was compiled in China in the 17th century and takes 1500 volumes in the modern reprint.
2. More information on this project can be found at `<http://coe21.zinbun.kyoto-u.ac.jp/>` the website of the Institute for Research in Humanities, COE 21 section
3. As defined in ISO 13250 (International Organization for Standardization)
4. Most of this had been done semi-automatically. Just to give an idea of the amount of the material, the size of the files with only very basic encoding applied runs at this moment to well above 30 MB.

Bibliography

ISO. *ISO/IEC 13250, Information technology - SGML Applications - Topic Maps.* Geneva: ISO, 2000.

Liu Xu. *舊唐書 (Jiu Tang Shu) (945).* Beijing: Zhonghua Shuju, 1975.

Ouyang Xiu. *新唐書 (Xin Tang Shu) (1060).* Beijing: Zhonghua Shuju, 1975.

Sima Guang. *資治通鑑 (Zizhi Tongjian) (1084).* Beijing: Zhonghua Shuju, 1956.

Wenyuange Siku quanshu dianziban. Hong Kong: Chinese University Press, 1998.