**HO — Markus**

<2 figures>

**MARKUS: An Infrastructural Semi-Automatic Markup Platform for Classical Chinese**

Ho, H. I.

The approach of digital humanities has widely interested many humanists from all disciplines. We can see it in the new methodologies introduced at DH2014 in Lausanne, where more than 700 registered participants gathered from around the world. However, digital humanities are still new practices and, in many cases, unachievable practices for many humanists. In the international workshop ‘New Perspective on Comparative Medieval History: China and Europe, 800–1600’, which took place in Oxford in January 2014, one discussion about ‘Isn’t the Siku Quanshu (Database) Enough?’1 reflected a common but critical debate between two groups of humanists. Scholars are satisfied with large commercial text databases, and they question why their colleagues invest their research time in data preparation (for example, encoding research texts in TEI) for computational analysis rather than read through the entire search results returned by databases. In the specific case of encoding texts in TEI, scholars often find themselves spending years doing manual encoding before computational analysis can be applied, despite the fact that the TEI standard has already saved a lot of work in schema design. We propose that in addition to defining a standard schema for encoders, efforts must be made to develop semi-automatic markup tools to speed up the tagging process.

In this paper we will introduce MARKUS, a semi-automatic markup platform that is designed to automate the markup process of different kinds of named entities in the domain of classical Chinese (historical) texts. In particular, we focus on the possibility of providing an infrastructural but low-cost and thus sustainable markup service that is solely built upon JavaScript and HTML5, both extremely basic and well-supported technologies.

MARKUS was developed as a tool to speed up the tagging process for the Communication and Empires project (http://chinese-empires.eu), which applied the TEI-markup approach to a corpus of 112 notebooks of the Song dynasty of historical China. We manually tagged quotes, interlocutors, authors, titles, and topics for each entry of five notebooks following the TEI standard. However, based on the above tagging experience, we realized that if we wanted to analyze all the people mentioned in the texts, it would be simply impossible to manually tag them within the limited time of the project; every hour we could only finish approximately six to seven tags manually. This labor-intensive tagging process is a common barrier for humanists researchers interested in putting the digital humanities approach into practice. Therefore, MARKUS aims to be an infrastructural, user-friendly, openly available, and sustainable markup service for Sinologists to overcome this barrier of encoding texts.

MARKUS currently provides three markup functions to help its users to tag classical Chinese texts: automated markup, keyword markup, and manual markup. Instead of providing a centralized and powerful (often complicated) web application, we try to make the service easy to operate by separating each markup function into different single web pages as a single task (Figure 1). In the meantime, all the web pages still share consistent interface design (Figure 2). Users can focus on a single task at a time while still following our step-by-step workflow to accomplish the entire tagging process.

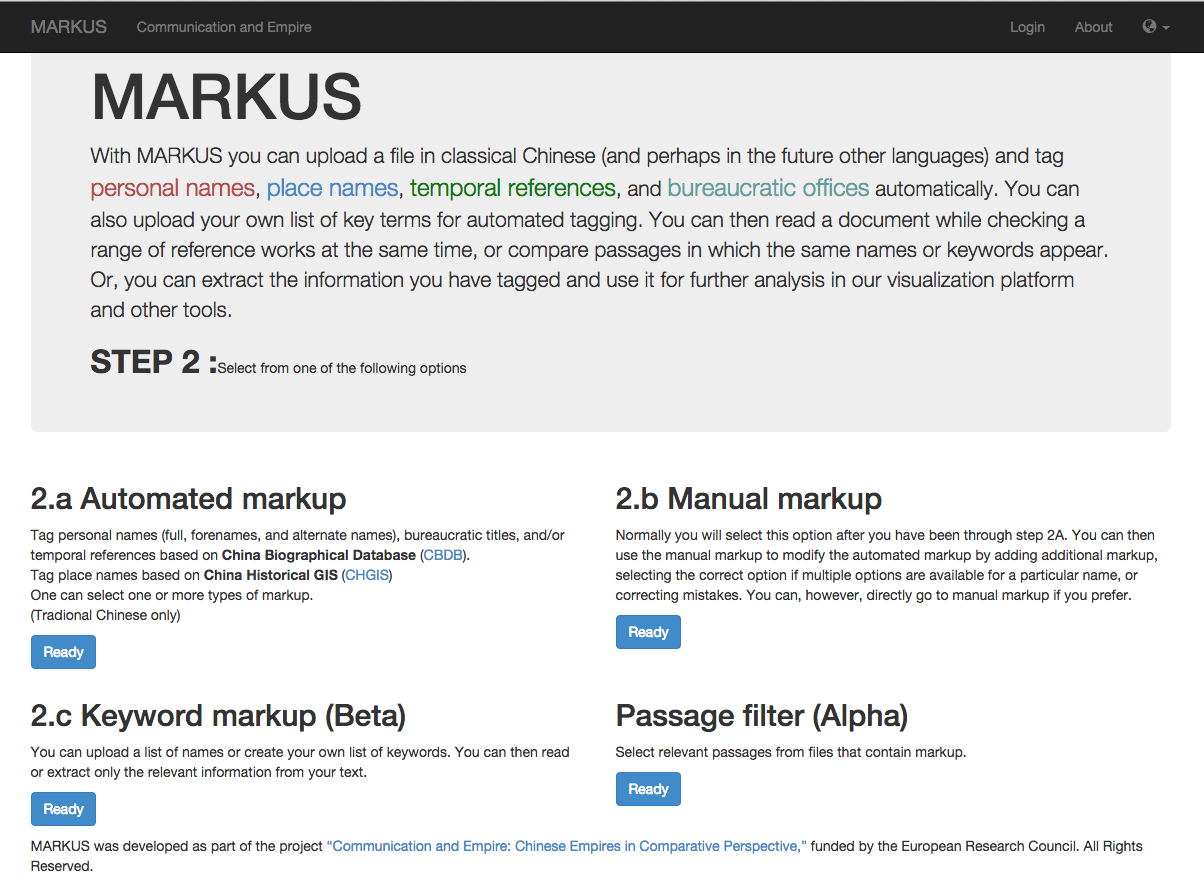


Figure 1. The step-by-step workflow interface.

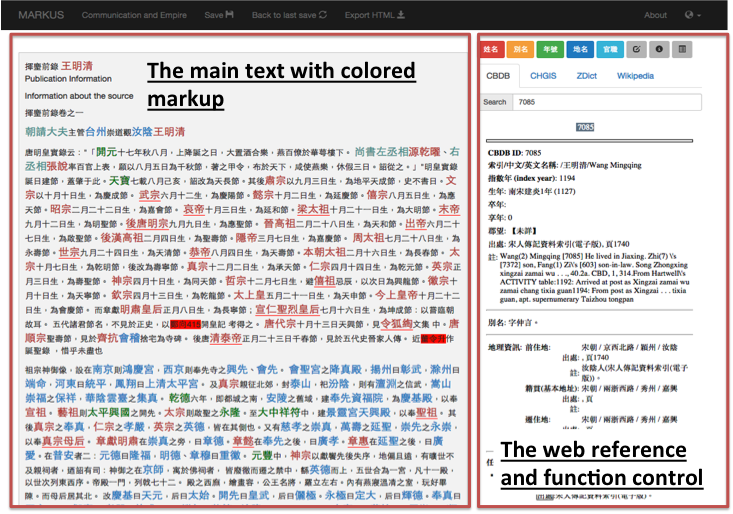


Figure 2. All markup functions follow a consistent interface design.

The workflow starts with uploading a text file to MARKUS (step 1). After the text is loaded, the user can use the automated markup function (step 2.a) to scan all named entities known to the system. Then the user can choose to apply keyword markup (step 2.b) to scan and tag texts against a list of terms or a regular expression given by the user. At the last step (step 2.c), the user can verify and refine all the markups manually.

The automated markup function of MARKUS is currently capable of identifying commonly needed types of named entities in Chinese historical research. MARKUS is built in with 355,000 personal names, place names, temporal references, and official titles based on the results of other digital projects, namely the China Biographical Database (CBDB; http://isites.harvard.edu/icb/icb.do?keyword=k16229) and the China Historical GIS (CHGIS; http://www.fas.harvard.edu/~chgis/). Named entities for more specific research interests—for example, terms collected in the Buddhist Studies Authority Database Project (http://authority.ddbc.edu.tw/)—will be incorporated.

MARKUS uses a color-coding scheme to display markups according to their tags. MARKUS also associates tagged texts with their sources by an identifier defined in the sources in order to provide better interoperability between projects. For example, MARKUS associates a personal name with its CBDB person ID, so that users can link to CBDB to get more information about the person.

The built-in lists of named entities only covers the basic needs of Sinologists. Alternatively, MARKUS allows users to upload their own lists of terms and to write regular expressions to tag texts with observable patterns. After the automated and/or the keyword markup, users can validate, add, or remove markups manually while reading through the text. A range of online reference tools and dictionaries (CBDB, CHGIS, ZDICT,2 and Wikipedia) have been integrated into the interface to assist the reading and validating process. We also provide a batch edit function to speed up the process of removing or adding a tag throughout the text in a batch. The markup result can be saved as a MARKUS file or exported as a HTML or a TEI file. Users can also further choose to export all the tags along with the passage identifiers to a tabular file (CSV, HTML table, or Excel file) to conduct further statistical, temporal, spatial, and social network analyses.

MARKUS is still in its early development stage. More functionality, such as visualizing the markups in charts, maps, or a timeline, will be added to provide an infrastructural markup service for Sinologists. MARKUS is freely accessible via http://dh.chinese-empires.eu/beta, and the source code will be released at the end of the project.

In order to make MARKUS as sustainable as possible, we chose to develop MARKUS as a non-centralized web application with the most basic technologies. It is solely written in JavaScript and HTML5 without any server side scripts,3 meaning it can be hosted in any free web hosting service, even those with the most basic facilities. It is then quite possible to provide the MARKUS service for a long term without any funding. However, this also limits the computing power that MARKUS can provide. During the upcoming development phase, while MARKUS is still funded, we plan to extend MARKUS to include more advanced markup functionalities such as applying machine learning and text mining techniques for automatic markup. It will require a dedicated server to provide higher computing power and online storage for uploaded texts, which will lead MARKUS toward a server-side implementation.

**Funding**

This work was supported by the Arts & Humanities Research Council and the National Endowment for the Humanities.

**Notes**

1. Siku quanshu is a one of the largest digital corpora of classical Chinese texts for Chinese cultural studies. It has been digitized as a commercial database. The discussion has been depicted in a blog post by Hilde De Weerdt with her view on this common debate (De Weerdt, 2014).

2. ZDIC (http://www.zdic.net/) is an online Chinese dictionary.

3. MARKUS requires HTML5 web worker (http://www.w3.org/TR/workers/) and File (http://www.w3.org/TR/FileAPI/) API to provide a better user experience in heavy computation process and file loading/saving functions. JQuery (http://jquery.com) and Bootstrap (http://getbootstrap.com) are the major JavaScript libraries used in MARKUS.

**Reference**

**De Weerdt, H.** (2014). Isn’t the Siku Quanshu Enough? Reflections on the Impact of New Digital Tools for Classical Chinese. http://chinese-empires.eu/blog/isnt-the-siku-quanshu-enough-reflections-on-the-impact-of-new-digital-tools-for-classical-chinese/.