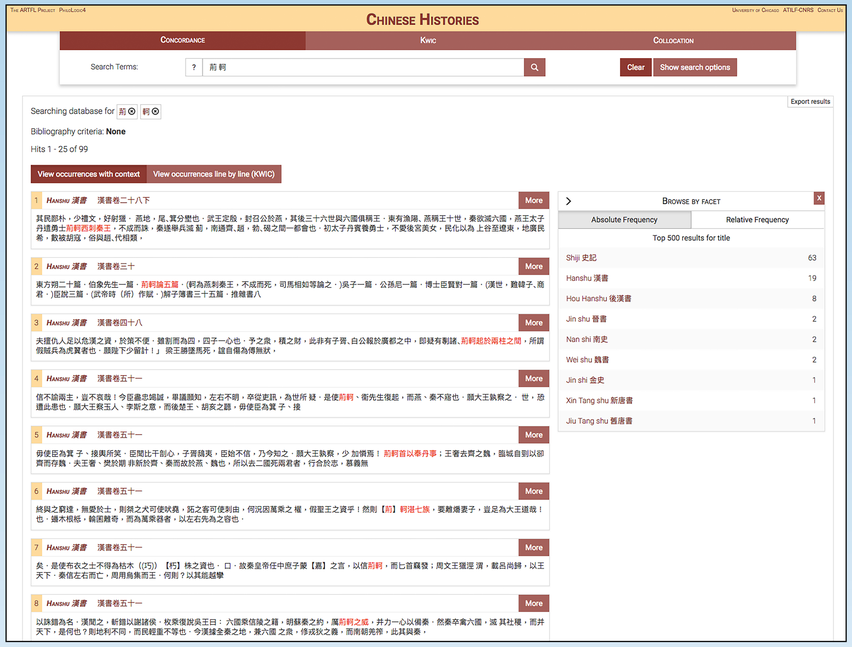
Textual Optics: Converting the 24 Chinese Histories to PhiloLogic

PIs: Robert Morrissey (Romance Languages), Hoyt Long (EALC), Haun Saussy (EALC/CompLit), James Sparrow (History), Clovis Gladstone (ARTFL), Jeffrey Tharsen (RCC)

PhiloLogic and Large-Scale Intertextual Analyses

PhiloLogic was originally developed in 2001 by the ARTFL-FRANTEXT group at the University of Chicago under the direction of Robert Morrissey. Initially designed for French corpora, at present there are over 100 textual databases under PhiloLogic, in languages such as ancient Greek, Latin, Hindi and Urdu as well as nearly all Western European languages.

The development of the Aozora Bunko 青空文庫 Japanese textual corpus for PhiloLogic in 2015 and the refinements to the system that allow for searching of non-space-delimited languages has now permitted the inclusion of Chinese-language corpora. The first two large-scale Chinese corpora to be brought into PhiloLogic are the "Republican China Periodicals" corpus and the "24 Chinese Histories" corpus comprised of all of the official histories produced by court historians over the past 2,200 years; the Histories corpus further includes searchable "Notes" fields with the tens of thousands of traditional commentarial exegeses from the Zhonghua Shuju 中華書局 editions of the texts.



The PhiloLogic Online Search Interface

Via the PhiloLogic web-based Search Interface, users can find any specific words or terms or combinations of words and/or terms (including word stems and wildcards like "*" to represent unknown or variable characters), and see immediately where they occur (or co-occur) in any of the texts in the database. The simple example above shows the name of Jing Ke 荊軻, a famous assassin, and the contexts in which his name and the story of his attempt to assassinate the first emperor of China occur in the Histories.

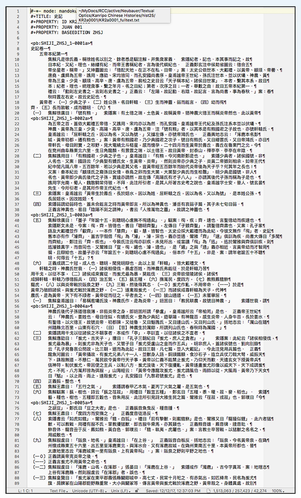
Converting the Kanseki Repository texts to TEI

As PhiloLogic requires tagged TEI-formatted texts, the main hurdle we needed to overcome was how one could convert the proofed versions of the source files from in Kyoto University's Kanseki Repository (www.kanripo.org) to PhiloLogic TEI format.

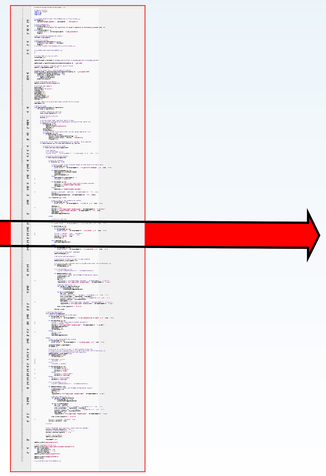
The 320-line Python script (below center) developed by the RCC allowed the formatting of various textual elements and compositional structures to be detected algorithmically (the original source is shown in the lower left-hand corner), and then these sections were concatenated, auto-tagged with custom TEI tags and reconstituted in the format required by PhiloLogic (as shown in the lower right).

All inline commentarial notes were replaced with unique "notes" pointers so that users can view them in pop-up windows while reading the head text, and custom metadata fields were created to allow for constraining of the search results based upon a range of detectable criteria within the texts (e.g. restricting search results to specific sections or chapters, types of entries, or sorted chronologically).

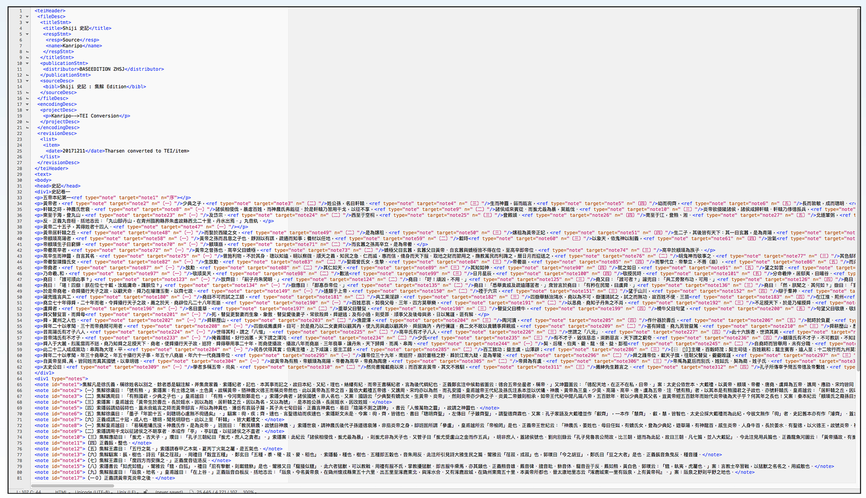
As the Kanseki Repository uses a standard file format for the over 9,000 individual texts in their archive, the development of a custom conversion algorithm means that we can now include any or all of the works from that archive (or any others that use their format) in the PhiloLogic4 system here at the University of Chicago. We look forward to continuing to partner with Kyoto University and Christian Wittern, Kyoto's digital library holdings specialist, as we begin to incorporate more and more diverse types of Chinese texts into PhiloLogic.



Original source text



Python Conversion Algorithm



New PhiloLogic4-compliant TEI-tagged version of the source