# Prospects for the Confucian Etext Project (儒家電文計劃)

## Stephen C. Angle, Wesleyan University, USA

### Introduction

It is by now a cliché that our future will be closely tied to Asia. Our success in this future will depend to no small degree on how well we understand Asian cultural traditions, for with better understanding comes better and more fruitful communication. The goal of this project is to provide a resource for students and scholars that will dramatically enhance our ability to access and interpret the core philosophical texts from China's Confucian tradition. The Confucian Etext Project aims to provide Internet access to and sophisticated searching of the full Chinese texts of major works by the most important Neo-Confucian thinkers, as well as those of the four most influential Confucian-inspired writers from the end of the Imperial era.

We now have ample evidence that access to philosophical texts in electronic form can revolutionize the possibility for research and understanding in an academic field. A decade ago the Thesaurus Linguae Graecae (TLG**)**, a CD-Rom containing large numbers of Greek classical texts in digital form, dramatically reduced the amount of time necessary for combing through texts to see how and where authors used particular terms or phrases. Research that would have taken months now took hours or even minutes, making possible projects that would never have been practical before.

More recently, the Perseus Project < http://www.perseus.org > has demonstrated the advantages of (1) making text databases available over the web, and (2) linking texts to lexicographic information. The combination of these two steps has taken Perseus beyond purely research applications -- though it remains a superb research tool -- and into the realm of pedagogy. Perseus can now be used as part of a language course or even as part of literature or history courses, so long as students have at least minimal background in the language, since the integrated lexicon allows readers at all levels to work with texts in their original language. In fact Perseus has an audience even beyond current students: anyone who has ever studied Greek can, for a minimal fee, access the database and read the classics with the help of the lexicon.

Students and researchers working on Chinese philosophical texts currently face even larger barriers than our counterparts in Greek classical studies did prior to the TLG. Even before the Greek texts were digitized, there existed concordances and indexes that facilitated manual searches. For the vast bulk of the Confucian tradition, this is not the case. Concordances do exist for many Classical-era Confucian texts (pre-221 B.C.E.), and some of these texts have also been digitized.

Very little attention has been paid to Confucian texts after the classical era, however, leaving scholars in a close to untenable position. Gone are the days when Confucian scholars memorized great swathes of texts, but a remnant of those days is the lack of even simple indexes to the great philosophical classics from the eleventh century through the nineteenth. If a student today wants to investigate what Zhu Xi 朱義 (1130-1200) said about a particular term or concept, he or she has but one choice: read through thousands of pages looking for the term or concept to appear. This is all the more true since Chinese philosophical writing differs significantly from Western versions of the genre: thoughts and conversations tend to be recorded chronologically, with only the loosest topical organization applied, if even that.

**Current Situation**

To date, only two groups have recognized and attempted to improve this situation.

**A. One is my Confucian Etext Project <http://www.wesleyan.edu/~sangle/etext>**

The project began in the summer of 1994. Phase I of the project will last until the end of the upcoming academic year (May 1999). It has been funded largely by internal seed money, supplemented by a small grants from the New England Council of the Association for Asian Studies, the Mansfield Freeman Center for East Asian Studies, and most recently the Keck Foundation. These monies have provided for hardware and software acquisition and for salaries for student research assistants. Phase I has had three goals:

(1) Develop proceudres for accurate and efficient preparation of texts.

My interest in the project began when optical character recognition (OCR) software for Chinese became available. All of the texts thusfar prepared by the Project have been input using the best of the available OCR packages, "Dan Qing Professional 丹青專業板" by UMAX.

While our results with OCR have been acceptable, it has now become clear that contracting to have employees at Chinese research institutions prepare the texts through double-typing is an even better alternative: it provides both a better value and more accuracy.

(2) Prepare initial texts, chosen by criteria discussed below, format using HTML, and distribute on the World Wide Web. Receive and evaluate feedback on ease of access and use.

Use (as measured anecdotally and from web logs) has been significant and ever-increasing. Some of our earliest texts are less than ideally accurate, and some method of rating "authoritativeness" has

been asked for.

(3) Explore options for more advanced tagging of texts and, in cooperation with the Wesleyan University Library, test and acquire an object-oriented fulltext search engine that is compatible with Chinese text.

Phase I of the project has so far only scratched the surface of what a well-designed text database can be, since we have not yet implemented sophisticated tagging or advanced searching. We are preparing to do so, however, and this will be a central part of Phase II. Beginning in the summer of 1998, we have funding to begin preliminary tagging with the TEI Lite DTD. We will also experiment with lexicon integration and sophisticated searching. Our goal will be to be fully prepared to make Phase II a reality beginning summer.

**B. The second is the Institute of History and Philology at Taiwan's Academia Sinica**

This group has assembled a large corpus of texts <http://www.sinica.edu.tw/ftms-bin/ftmsw3>. To date, this project differs from my own in several significant ways.

(1) Its scope, personnel, and funding are greater.

(2) It concentrates on classical-era works, especially for philosophical texts. There is very little overlap between its texts and our own.

(3) Only certain institutions who have entered into expensive contractual agreements with the Institute have access to all the texts.

(4) Access, whether free or purchased, is limited to simple term-searching; also, the texts themselves cannot be downloaded.

(5) The texts are not integrated with a lexicon nor linked in any ways with one another.

In each case there are good reasons for these decisions; my goal today is not to criticize these choices, but to explore alternatives and open up a discussion about the range of resources scholars and students need.

**Future**

**Selection of Texts.** Our over-all scope includes Confucian and Confucian-inspired texts from the

11th century C. E. to the present. This covers a vast number of texts, at least comparable to what one would find in Latin and vernacular Western language if one were to look at "Christian and Christian-inspired texts" over a similar period. Many of these texts are absolutely fundamental to the shape Chinese culture took over this last millennium, and continue to be important as the new millennium begins. Given this vast potential corpus, we have used the following criteria in establishing which texts, in which order, we prepare:

- Does a scholarly adequate edition of the text available in digital form elsewhere?
- If so, does the community of students and scholars has access to the text(s)?
- Is the text widely recognized as philosophically significant?

**Input Procedures.** In Phase I of the Project we used OCR as our primary input method, dealt with only one version of a given text, and did not incorporate tagging. In Phase II, our procedure will be as follows:

(1) One edition of a given text is photocopied and sent to our text input center.

(2) There the text is typed twice and the two versions are compared against one another in an initial proofreading process. Then the text is proofread again. Like participants in the Buddhist Text Initiative <http://www.acmuller.gol.com/ebti.htm > who have worked with such groups, our contract will specify that the error rate cannot exceed one mistake per one thousand characters. When the work is completed, the final version is mailed (or emailed) back to Wesleyan.

(3) Native Chinese-speaking student research assistants compare the text against a second (different) edition of the text, and note variant readings. This also serves as an additional proofreading stage.

(4) The text is tagged, and links to the lexicon generated automatically.

(5) In the end we will have two versions of each text, one in SGML format for downloading, and one in KE Texpress format for on-line reading and searching. The base text will be the SGML version, using the TEI Lite DTD; we will then run that text through a filter program to load the database into KE Texpress for on-line reading and searching.

**Markup of Texts.** Phase I has involved only minimal markup of the texts: we have used HTML to indicate basic formatting. In Phase II, we will use SGML (or XML) as our mark-up language. The benefits of SGML in the production of scholarly editions are well-known and well documented. The particular DTD (document type definition) we will use will be TEI Lite (see <http://www-

tei.uic.edu/orgs/tei/>), which contains all the tags that we will require and is considerably simpler to implement than earlier DTDs.

Much of the tagging is routine work that can readily be automated, as for instance the individuation of paragraphs and sentences. Other work requires textual analysis and will be done by the Project Manager under the guidance of the Project Director. Our scope includes:

- structure (e.g., chapters and sections) and
- content (e.g., people, books, and quotations), and
- editorial apparatus (e.g., differences between editions),

but does **not** include

- most grammatical information (e.g., part of speech).

This limitation stems both from the interests of the Project Director, as well as from the extreme difficulty of automatically tagging grammatical categories in a language, like Classical Chinese, which is not inflected. Manually tagging the part of speech of each of the hundreds of thousands of words in our corpus is impractical.

**Lexicon Integration.** The Classical Chinese lexicon we will use will be based on Charles Muller's WWW CJK English Dictionary Database <http://www.acmuller.gol.com/cjkdict.htm>. Dr. Muller notes that "The number of single characters with full information contained in the CJK Dictionary already exceeds that of *Mathews*," one of the current standard printed references for Classical Chinese. We will write a filter program to load the Lexicon into KE Texpress. We will also write scripts, in Java or PERL, to create the dynamic links between the text and lexicon.

Once the dictionary has been modified for our needs--a project with which Dr. Muller has expressed great interest in participating--we will establish links between every character in our texts and their dictionary entries. This is a simple process that can be performed automatically, since the Big-5 or Unicode code for each character is tagged as part of the corresponding dictionary entry. We will also link every compound contained in both the dictionary and one or more of our texts to its corresponding entry. Compounds are quite infrequent in Classical Chinese, but this process, when combined with tagging of people, places, and book titles, will further enhance a reader's ability to make sense of these texts. Once implemented, clicking on a character (or compound, or proper name) will bring up a new window on one's screen containing the relevant dictionary information.

**Search Engine.** Once the texts have been marked up using SGML, they can be immediately

indexed for use with the object-oriented fulltext search engine the Wesleyan University Library has acquired, KE Texpress <http://www.kesoft.com>. The library selected KE Software for a variety of reasons:

- it is among the fastest of full-text search engines;
- it can handle SGML as well as other marked up language databases;
- it supports multi-lingual data, including Chinese 2-byte characters;
- it is very cost effective; and
- the company has the best customer support among companies investigated.

Using KE Texpress, our library has successfully setup two SGML databases on the web, the *Oxford English Dictionary* and *Chadwyck-Healy's English Poetry*.

KE Texpress is a object-oriented database management program. Since it can not interpret DTDs automatically, we will write a filter program to load the final text into KE Texpress. We will also create the dynamic links between the text and Lexicon with Java or PERL scripts as well as modify and refine the searching interface.

**Editorial Procedures.** A crucial issue in preparing a text database is accuracy. As outlined above, our procedures call for two different editions to be compared with one another. Many differences between the two will be simple (and obvious) input errors, which will be immediately corrected; where differences stem form differences in the editions themselves, the differences will be fully preserved in our edition using specialized SGML tags. The Project Manager will largely be responsible for this process, though always under the oversight of the Project Director, who bears sole responsibility for and final control over the content of the database. To assist the Project Director when necessary, we have established an advisory Board made up of both senior scholars who use the texts and technical specialists familiar with large-scale text-encoding projects, from whom we will solicit regular input on possible improvements.

**Presentation.** All our materials will be presented to the public via the World Wide Web. A large and ever-increasing percentage of the student and scholar audience of the project has ready access to the web, making it an ideal medium for disseminating the material.

We plan to offer both downloadable full text files of each text and access to the texts via the search engine described above. It is important to enable both the casual browser of the database to easily access the materials, and to provide tools for scholars to make more advanced use of the texts. In addition, allowing users to download SGML-tagged texts enables the most sophisticated users to use tools of their own choosing, or their own design, with our texts.

**Issues**

**(1) Copyright.** Since for each text, we will be preparing our own edition (relying on at least two existing editions, as well as considerable additional research, throughout the input and tagging process), we will own the copyright to our texts and can distribute them as we see fit.

**(2) Multiple formats.** The first method of access is to use the texts directly on our web site. This allows users to search the texts with the KE Texpress search engine and to utilize the integrated Classical Chinese lexicon. The second method of access is to download the full texts onto their own computers. We are committed to supporting this option as we recognize its value to many users. We are also considering a third method of access, namely reading the SGML texts directly on the web. One option is the free browser plug-in SoftQuad Panorama, which will enable a standard HTML browser to properly display SGML data. If users simply want to browse (rather than search) the texts and use of the integrated lecixon, this may be the method of choice.