Eighteenth International Unicode Conference (IUC18)
Unicode and the Web: the Global Connection
April 24-27, 2001, Hong Kong
http://www.unicode.org/iuc/iuc18/

SUBMISSION ABSTRACT

• PRESENTER:

Richard S. COOK
STEDT Project, Linguistics Department
University of California, Berkeley
mailto:rscook@socrates.berkeley.edu
http://stedt.berkeley.edu/

• PAPER TITLE:

The Extreme of Typographic Complexity:
Character Set Issues Relating to Computerization of
The Eastern Han Chinese Lexicon《説文解字》*Shuowenjiezi*

• STATEMENT OF PURPOSE:

This presentation is concerned with character set issues relating to computerization of one of the most important and most typographically complex Chinese texts,《説文解字》 *Shuowenjiezi* (SW). The title of the SW lexicon has been translated as 'Interpreting the Ancient Pictographs, Analyzing the Semantic-Phonetic Compounds' (Cook 1996). This Eastern Han Dynasty (121 A.D.) text was the first attempt at a systematic componential analysis of all of the characters in the complex Chinese writing system. With regard to this text, this paper addresses the following four topics, listed here, and briefly described below:

1.) The SW text -- its history, character and importance.
2.) The character forms -- their styles and components.
3.) The font -- the character set and production process.
4.) Encoding Standards -- mappings and missing characters.

• PAPER DESCRIPTION:

The paper begins with a brief introduction to the SW text, including its basic history, general characteristics, and overall importance to linguists, paleographers, epigraphers, and classicists. In particular, the linguistic importance of computerization of this text is emphasized.

The character forms found in the text are then discussed, with reference to both stylistic and componential issues. Special emphasis is given to the relationship between the text's componential analyses and the actual items of the character set. The issue of natural (extrapolated) extensions to the character set is mentioned.

Next, the 11,246 character font developed to capture this text is introduced. This is a CIDFont with Type 1 outlines. The rigors of the font production process are described, including hardware, software and indexing issues. Demonstration will be given of the typographic and lexicographic database systems employed in and resulting from the production process.

Finally and most prominently, encoding issues are addressed. Primary focus is given to mappings of the text-based character set to both Big-5 and Unicode standards. In this regard, mapping and missing character issues are discussed with illustrative examples.