Typological Encoding of Chinese: Characters, Not Glyphs

Richard S. COOK
STEDT Project
Linguistics Department
University of California, Berkeley
<rscook@socrates.berkeley.edu>
<http://stedt.berkeley.edu/>
2001/08/10/22:58

• Synopsis

This presentation is concerned with examining the relation between the following two points:

1.) The difficulties of holding the Chinese-related (CJKV) scripts answerable to the same principles governing other script encodings.

2.) The questions surrounding adequate typologization of CJKV scripts, and their implications for the encoding of other scripts.

Most specifically, the ideas mentioned in this paper relate to specialized usages of relatively rare CJKV ideographs. Example data will be presented from ISO/IEC 10646-1 :2000(E), and from a relational implementation of the Unihan.txt (v. 3.1.1d1) data file. Illustrative examples will also be drawn from the most recent *Wenlin* (Bishop, 2001), and from my *Shuowen* database.

• CONTENTS

◊ 0.) Introduction

The enormous, even "open-ended" unified CJKV character set, encoding at present more than 71,000 characters, perfectly illustrates the limitations of the "Characters, Not Glyphs" distinction drawn in the Unicode Standard (v. 3.0, e.g. p. 13).

Unicode's CJKV encoding is at present typologized only insofar as it is unified, which is to say that a given encoded character serves to typify an abstract character form which may be realized (as glyph) in a particular script (Chinese, Japanese, Korean or Vietnamese) according to the stylistic traditions of that script.

CJKV typologization could however be taken much further, and in fact arguably should be taken much further, if CJKV scripts would be held to the same standards that other scripts (e.g. Arabic, Myanmar, Tibetan ...) have been.

How and why have the CJKV characters not been held to the same standards, and what problems does this present? How might Unicode implementers benefit from componential data being added to the Unihan database, and what might such data look like? How do the Ideographic Description Characters (IDC) and Ideographic Description Sequences (IDS) relate to this?

This presentation explores these questions, and in illustrating the componential nature of CJKV characters, demonstrates a method for determining (on the basis of the present CJKV character set) a base extensible component set for generating and encoding infinite CJKV characters. It is shown that this framework permits closer adherence to the spirit of the "Characters, Not Glyphs" distinction.

It is argued that adequate historical typologization of the Chinese script can only be accomplished with reference to specific texts and inscriptions, and this is what I mean when I refer to a "Text-based" or "Source-based" encoding. These encodings may also be termed "Contextual" in that they seek to document the historical context whence the glyph usage derives. The written sources of ancient Chinese are many and varied, and each offers stylistic peculiarities and mapping challenges. Oracle Bone Inscriptions, Bronze Inscriptions, Stone and Earthenware Inscriptions ... all of these contain vital historical information which only a typological system can address.

A typologization adequate for Chinese purposes would be rather simple in some respects. First, it should characterize aspects of glyph shape, and second, it should characterize aspects of glyph usage. One may however complicate and broaden this scheme in many ways, for example with script-specific issues of componential encoding. It seems that distinctions adequate for the handling of the Chinese script must have implications for the encoding of other scripts.

◊ 1.) THE CJKV SCRIPTS

The Chinese and Chinese-derived scripts have a history extending back at least 3500 years. The Chinese script is first attested in inscriptions dated to about 1500 BC. Writing in China developed with various levels of standardization and variation through the next 1200 years. In the face of longstanding divergent usages (competing local standards), the first stirrings of national standardization of the Chinese script are usually said to have occurred in about 200 BC, in the realm of the first Qin Emperor (c. 259-210 BC). Based on the standardizations imposed in that short-lived though influential dynasty, some 300 years later in the Eastern Han Dynasty the mother of all Chinese dictionaries, *Shuo Wen Jie Zi* (121 AD), attempted to catalogue, analyze, define and classify all of the 11,000 or so characters employed in the script up to that point.[1] This Eastern Han dictionary became a defacto standard, and so it is that Chinese-derived characters today are widely known as "Han" characters.

In subsequent centuries, provided with this firm lexicographic basis, the Chinese continued to catalog and to invent new characters. By the 12th century when the next great character dictionary *Guangyun* was complied, the number of characters collected had more than doubled, to 25,126.[2] Already at this time non-Chinese-speaking peoples all over Asia had fallen under the mystical spell of the Han script. As time went on, the Japanese, Koreans and Vietnamese all adopted, augmented and customized it to suit the needs of their own languages.[3]

In each of these locales local varieties or local usages of the "Han" script have developed, some over hundreds of years. Graphical differences emerged between the different scripts, so that a given character in one locale was no longer identified with the character to which it was historically related in another locale. What had once simply been stylistic variations in print or handwriting gradually crystallized into hard irreconcilable differences.

With the advent of the computer, encoding standards for each of these script areas developed independently, or with only limited mutual awareness, each capriciously and imperfectly cataloguing the characters used in the different regions.[4] The methods for collecting characters for encoding varied widely, and within a given writing system variant forms of a given character were often encoded separately without regard to the relation among the variants.

For as long as there have been computers that could do anything with Chinese-derived scripts, computer users have taken it for granted that some of the characters that they needed were not available, might never be available. They have been accustomed to being unable to easily exchange electronic documents between systems. Until recently, among the best that computers had to offer specialized Han script users was an encoding known as "Big 5", an encoding which in English terms might be likened to a keyboard without the five (relatively low-frequency) letters k, j, z, x and q, and with no upper case at all.[5]

---

[1]Computerization of this text in its various editions is the subject of my ongoing work, and was the subject of my IUC-18 paper. Please see this paper for further details on the character counts in *Shuowen*.

[2]This count derives from my work in progress to computerize several editions of this text.

[3]Even the Japanese Kana syllabaries ultimately derive from Han characters.

[4]For detailed discussion of the various Han-derived encodings of Asia, see LUNDE.

[5]See my IUC-18 paper for more on this analogy.

When digitization of local script variants or historical texts was undertaken, users and developers began to appreciate the full extent of the problem, and the inherent limitations of the standardization process. Even with regard to a single locale, for example, mainland China, the issues of achieving an adequate standard are daunting, to say the least. The problems have been so great that only in the last few years, in the dawning of the 21st century, when personal computational power has progressed to levels unimagined twenty years ago, has an international standard for the Han-derived scripts emerged.

This international standard is Unicode. The immense work done to produce this standard, undertaken by the Ideographic Rapporteur Group (IRG)[6], has pushed CJKV[7] computing to higher levels than many had ever hoped possible. With IRG's creation of "Extension B", 42,711 new codepoints were added to the Unicode Standard, so that it now encodes 70,207 unique "ideographs".[8]

◊ 2.) THE CHARACTER-GLYPH MODEL

The Character-Glyph Model (CGM) as outlined in *The Unicode Standard* Version 3.0 Section 2.2 "Unicode Design Principles" is the high ideal which serves as the conceptual basis for Unicode. Despite the desire for legible plain text, despite the problems with adhering to this ideal in a world of legacy encodings and compatibility characters, the CGM stands firm when it can, and gracefully bends when it cannot stand firm, and when it cannot bend, it simply bows out. The CGM is not tyrranically enforced, but is rather accommodating to historical practice.

Simply put, the CGM is about the distinction between "abstract" and "concrete". It attempts to draw a firm line between an abstract ideal "character" form on the one hand, and a concrete real "glyph" on the other.

The *abstract character* form can be conceived as a basic outline of what is most essential about a particular character, the sum of all things that make character "a" distinct from character "b" and distinct from all other characters. Each abstract character form is assigned a unique number.

The *concrete glyph* exists in the real world on paper and on computer screens, displayed and printed in different font faces, in different styles, and written in different handwritings. A thousand people may write the letter "a" in a thousand different ways, and these thousand variants would each be a glyph. It is their common idea of the character "a" which allows each person to recognize the other 999 glyphs as the same meaningful unit "a".

---

[6]<http://www.cse.cuhk.edu.hk/~irg/>

[7]Chinese, Japanese, Korean, Vietnamese.

[8]The term "ideograph" is a technical usage defined in the glossary of the Unicode Standard, a compromise term equivalent to "CJKV character".
For these counts, I am indebted to John Jenkins. Although in his IUC-18 presentation he claimed (jokingly) that "Every time I add up the totals I come up with a different figure.", in fact, he and the IRG do seem to have hard figures. The issue is somewhat complicated by things such as "compatibility characters which are not actually compatibility characters", and so I defer to them for the latest statistics.
27,484 : CJKUI, CJKUIA (p. 258 of the Standard 3.0)
27,496 : CJKUI, CJKUIA (including 12 compatibility ideographs that are not compatibility ideographs)
42,711 : CJKUIB (Extension B)
70,207 : total number of unique ideographs in Unicode 3.1

As an ideal, the Character-Glyph Model presents some serious challenges, not only to implementers at every level — from font designers to application designers to input method designers — but also to the average person seeking to understand what Unicode is all about and what advantages it may have to offer for their scrip, for their language and for their lives. For some scripts the constraints of the CGM present such a significant obstacle that developers may quite understandably shy away from Unicode implementation. A Myanmar reader might scan the Unicode code chart for his script and remark on the apparent fact that "some of our characters are missing". And so it is that the developer must step in to prove to the potential user that in fact nothing is missing at all, and that the designers of the Unicode Standard for Myanmar new exactly what they were doing, and did not make any mistakes at all in their design of that standard.

◊ 3.) THE CJKV IDEOGRAPHS MEET THE CHARACTER-GLYPH MODEL

Parallel to this, consider the Han-derived scripts discussed above. Until the advent of Unicode 3.1 Extension B, no developer would have tried to explain to a user seeking some missing Han character that the Unicode Standard had not in fact failed. Until Extension B, many needed (though rare) characters were still missing, and because of the design of the Standard, the best that could be done involved assignment of the needed characters to "private use" codepoints. In this regard, even with the considerably larger coverage of Unicode 3.0 in comparison with e.g. Big 5, certain specialized CJKV script users were still computing in a world where keyboards were missing keys.

Unicode 3.1 Extension B changed all this, empowering specialized CJKV script users as they had never been empowered before. With 70,207 unique ideographs currently encoded, one might imagine that every character ever to be needed had finally been assigned a codepoint. And one would be 99.999% right. In fact, the task of cataloguing and encoding that last tiny fraction of a percent of Chinese-derived characters is not complete, and on the contrary, shows signs of never being completed.

A recent report from US representatives to the last IRG meeting in Hong Kong illustrates the current state of things, with a rough breakdown per national body of ~67,000 proposed additions.[9]

ROK	23000+~20000
TCA	18000
PRC	4570
Japan	~200
Macau	~200
Vietnam	1049
HKSAR	9
DPRK	94

It seems that the great bulk of the additions being proposed for Extension C (and perhaps beyond) relate to work in the Republic of Korea on a Buddhist text called the *Tripitaka*.

If the idea of a Unicode Standard 3.1 with 70,207 unique ideographs doesn't surprise you, then I suppose you will also not think twice about a number almost twice as large. And yet, even a number that large will fail to permanently satisfy non-specialist and specialist users and scholars. And the reasons for this, complex as they are, relate primarily to non-adherence to the CGM, in the encoding of ideographic glyphs (variants) rather than characters.

[9]This data provided by Hideki Hiura, Sun Microsystems.

◊ 4.) UNIFICATION

To examine the way in which the CGM has not been followed, let's first examine what Unihan CJKV "unification" is all about.

Unicode's CJKV encoding is at present typologized only insofar as it is unified, which is to say that a given encoded character serves to typify an abstract character form which may be realized (as glyph) in a particular script (Chinese, Japanese, Korean or Vietnamese) according to the stylistic traditions of that script.

Looking, for example, at ISO/IEC 10646-1:2000(E) one can get an idea of what "Unification" is all about. In short, beginning at page 305, this document tabulates 5 of the variant glyphs unified under a particular codepoint. Each of the 5 glyph forms is classified under one of 4 larger headings: CJKV (the C source being split into G=Mainland China and T= Taiwan). Within each of these 5 major classes GTJKV are listed the contributing coded character set standards. A fuller list of the sources would be G, T, H, J, K, KP, and V, since Hong Kong (H) and the Democratic People's Republic of Korea (KP) now have sources.[10]

Unification in the *Unicode Standard* is in this sense "horizontal", in that glyphs have been classified together based on the similarity (if not actual identity) of the forms submitted by the various national bodies. The question of how similarity is gaged is a large one, which is mentioned further below.

Unification has not however been "vertical". What this means is that a given form (let's call it a glyph here) may actually be assigned more than one codepoint in Unicode. This occurs sometimes because of compatibility issues, but also sometimes because of simple failure to unify variants. For example, in a perfect world there are regular transformations which would allow for the "simplification" of a "traditional" character form. If a codepoint U+XXX was encoded which simply meant "perform the regular simplification transformations until I tell you to stop", then one aspect of the problem of unencoded simplified (Mainland Chinese) characters would simply go away. Such unencoded characters remain a problem for the very reason that simplification is a regular productive process, and the number of traditional characters greatly outnumbers the total number of simplified characters. This would provide a means of limiting the future bulk of the standard, which might be required to encode both simplified and traditional Chinese characters separately, even for extremely rare characters.

Beyond such a simple example, Extension B introduces many many more variant character forms which in fact ought to have been identified as variants of characters which were already encoded. These encoded variants are of several kinds, including "etymological" forms created to represent inscriptional Chinese of different historical periods. Solutions to variant selection (VS) are still being explored, blurring the line between plain text and markup/metadata, and so it is no surprise that without this mechanism firmly in place the IRG simply chose to encode new characters. "Horizontal" variant selection (HVS) is a relatively simple matter: e.g., when in Japan, use a

[10]Unicode 3.0 lists also a U source which has not yet been acted upon by the IRG. According to Jenkins, "the U source consists of U+FA0E, U+FA0F U+FA11, U+FA13, U+FA14, U+FA1F, U+FA21, U+FA23, U+FA24, U+FA27, U+FA28, and U+FA29, twelve characters which were put in the CJK Compatibility Ideograph block because they were derived from coroporate and not national standards, but which are not actually unifiable with anything else in Unihan." (personal communication, 28 Jun 2001) In brief, compatibility characters involve "round-trip" mapping between different standards.

Japanese font. "Vertical" variant selection (VVS) is however a much more complicated thing, especially without a clear means of quantifying sameness and difference.

So, unification represents one example of the CJKV character set conforming to the CGM, though we have also seen examples in which the CGM has been ignored.

◊ 5.) THE COMPONENTIAL STRUCTURE OF CJKV CHARACTERS

Another instance of non-adherence to the CGM has graver implications for the IRG's character repertoire, and this relates to the componential nature of Chinese-derived characters. It is well known that Chinese characters are not all simply unrelated units, but that all characters in fact fall into two broad classes.

- The first class of characters may be termed "Graphical Primitives", which is to say that they are not composed by the regular conjoining of other elements.
- The second class of characters may be termed "Compound", and comprises the class of all characters which are formed by the conjunction of simpler elements.

The situation is much the same as with English spelling. All words are formed by combinations of the fundamental 26 letters of the alphabet, which with case transformations total 52. If English words were encoded without regard to alphabetic decomposition, the situation would be every bit as bad as it is for the encoding of CJKV characters. A Unicode standard for English orthography without properly typologized component data is like a Unicode that starts encoding every mis-spelling and variant spelling of every word in the history of the English language. Clearly, many thousands of words would have to be encoded before it became possible to write all but the most simple English sentences.

But the analogy here is not quite perfect. Although it is a fairly simple matter to identify the elemental units of English orthography, it is much less easy to identify them for Chinese. To do this kind of identification properly, first an inventory of "all" the characters in the script must be produced, and this inventory would look very much like the CJKV portion of Unicode 3.1.

Once this character collection has been made, it then becomes necessary to analyze each character in a typological component framework, which is to say that components must themselves be grouped together into character classes of glyphs.

Two broad classes of component may be identified.

- One is the "etymological" component, based on the traditional understanding of the character derivation (often of a simpler form from a more complex form).
- The other is the "graphic" component analysis based on the actual shape of the character as it is written now.

The two analyses will in many cases be the same, and where they are not, the information has larger typological significance.

As discussed in my IUC-18 presentation (Hong Kong, 2001), I have been working for some time to digitize several editions of the Eastern Han Dynasty dictionary *Shuo Wen Jie Zi*, and it is on the basis of this computerization work that much of the traditional etymological componential analysis of the core script elements becomes available for the first time in Unicode compliant electronic format. Systems for componential analysis of Chinese characters have also been under

development by Wenlin Software[11] for quite some time now, and their pioneering systems seem nothing short of amazing, to say the least.

In a properly typologized componential framework, it becomes possible to encode a relatively small component set, with which all compound characters may be generated. In such a framework it does not matter whether a character has ever been encoded before. A user can encode a needed character as it is needed. And within the limits of the system, two users working independently in such a framework could encode the same new character in the same way. This would be similar to the situation in which two English typists have to spell a new word that neither of them has seen or heard before. Armed only with basic knowledge of English orthographic rules, they could come up with identical spellings. The problem is that where the English typists must fall back on known orthographic conventions, the Chinese typists must rely on the limits imposed by the computing system.

Answers to the question of "What might be a properly typologized componential framework?" must necessarily have some idiosyncracies. It is certain, however, that on the firm basis of the traditional Eastern Han analyses an adequate system can be devised.

◊ 6.) CONCLUSIONS

For my own purposes as a student of the history of the Chinese language, I am gratified that the Unicode Standard adheres as closely as it does to the CGM, but would of course be happier to see closer adherence. It is in fact difficult for me to conceive of a standard as being adequate which does not address the issue of graphic variation within a mechanism for etymological analysis. Such an ideal and admittedly complex mechanism is clearly at odds in some ways with the practical limits and purpose of an international data encoding standard, and yet the paradox seems to be that eventual emergence of such a mechanism will be the only completely satisfying long-term solution to the practical problems of international CJKV encoding issues.

The mechanism envisaged here involves specific bibliographic sources. It is a scheme like VSx (the variation selector scheme) in which the value of VS points to precise bibliographic info on the source, and x (another class of VS) serves to uniquely identify the character indexed within that source. Given VSS (variant source selector) and GID (glyph identifier), the sequence VSS + GID uniquely identifies a particular glyph within a particular source.

How do we measure glyphic variation? This can be done with typologized componential analysis, down to the stroke level, if necessary. Given adequate component data, "graphical distance" can be assigned a numeric quantity. If the character/glyph distinction is not black & white, but more as if glyphs are points along a spectrum of variation, what kind of encoding model would be best suited to this? In the spectrum (continuum) model, every character is a variant of every other character, via the continuum of attested glyphs. A conscientiously applied source-based componential typologization can handle this broad definition of "character".

[11] <http://www.wenlin.com/>

APPENDIX 1: ABBREVIATIONS AND GLOSSARY

ALSL: Apple LiSung Light, my reference Big5 font (cf. CLK below).
ATM: Adobe Type Manager; <http://www.adobe.com/>.
Big5: A common Traditional Chinese double-byte encoding standard (cf. Lunde 1999:171).
Bushou: 部首 Chinese 'lexical classifiers', a.k.a. "Radicals".
CGM: Character-Glyph Model. See the Unicode Standard 3.0, p. 13.
Chongwen: 重文 Chinese 'historical character stylistic variant'. See Section 2 above.
CIDFont: Character I. D. Font. Cf. Lunde (1999, p.288*ff*).
CLK: Apple Computer's *Chinese Language Kit*, a MacOS 9 install option (see Appendix 3).
CMap: Character Map. Cf. Lunde (1999, p.290*ff*).
Ext. B: Unicode 3.1's addition of 42,711 new ideographic codepoints.
FMP: FileMaker Pro (see Appendix 3).
FOG: Macromedia Fontographer.
GIF, JPG, **JPEG**: Digital graphical image file formats.
GSR: Cf. KARLGREN in the References.
Han: 漢 'Chinese dynasty; 西漢 Western Han (206 B.C.-24 A.D.), 東漢 Eastern Han (25-220).
HY: Harvard Yan Jing (Yen-Ching) Library, <http://hcl.harvard.edu/harvard-yenching/>.
HYDZD:《漢語大字典》*Hanyu Da Zidian.* Character lexicon. See 許力以 XU Liyi (1993).
HYPY: 漢語拼音 Hanyupinyin romanization of Modern Standard (Beijing) Chinese.
Jinwen: 金文 Chinese 'bronze inscription character'. Cf., e.g. 周法高 ZHOU Fagao (1981).
Kaishu: 楷書 Chinese 'square script', stylistic basis for modern Chinese typographic styles.
Kangxi:《康熙》 Character lexicon. Cf. 長玉書 ZHANG Yushu (1716).
LTBA: Journal *Linguistics of the Tibeto-Burman Area*, <http://stedt.berkeley.edu/ltba/>.
MacOS: Apple Macintosh Operating System. <http://www.apple.com/>.
MV: Main/Variant (cf. Section 4).
PDF: Portable Document Format; <http://www.adobe.com/>.
PPI: pixels per inch, computer monitor resolution.
Qing: 清 'Chinese dynasty name', (1644-1911).
Seal: 篆 Chinese character style. Cf. Xiaozhuan (see Section 2 above).
Shang: 商 'Chinese dynasty name', (~16th cent. B.C.-~1045 B.C.)
Song: 宋 'Chinese dynasties'; 北宋 Northern (960-1127) and 南宋 Southern (1127-1279).
Songti: 宋體 Chinese calligraphic style, cf. Kaishu.
STEDT: Sino-Tibetan Etymological Dictionary and Thesaurus, <http://stedt.berkeley.edu/>.
SW:《説文》。*Shuowen.* (Short for SWJZ.)
SWJZ:《説文解字》。*Shuowen Jiezi* (cf. References); also, the name of my BNU-based font.
SWJZ-FJZ:《説文解字・附檢字》。*Shuowen Jiezi - Fu Jianzi*
SWJZZ:《説文解字注》。*Shuowen Jiezi - Zhu* (cf. References); also the name of my font.
TIFF: Tagged Image File Format. A graphical image file format.
TTF: TrueType Font, a scalable font format, developed jointly by Apple and Microsoft.
Type 1: Adobe Systems' format for describing scalable fonts.
UCBEAL: UC Berkeley, East Asian Library, <http://www.lib.berkeley.edu/EAL/>.
Xiaozhuan: 小篆 Chinese 'Small-Seal character' (see Section 2 above).
XU Kai: 徐鍇 (920-974) 'Author of a Song version of SW', (cf. Appendix 6).
XU Shen: 許慎 (58?-147?) 'Eastern Han author of SW', (cf. Appendix 6).
XU Xuan: 徐鉉 (916-991) 'Author of a Song version of SW', (cf. Appendix 6).
Zhou: 周 'Dynasties'; 西周 W. Zhou (~1027-~771 BC), 東周 E. Zhou (~770-~256 BC).

APPENDIX 2: REFERENCES (SELECTED)

BISHOP, Thomas Eugene 畢曉普
2001 文林 *Wenlin.* Portland: Wenlin Institute. <http://www.wenlin.com/>.

COOK, Richard Sterling 曲理察
2001 "The Extreme of Typographic Complexity: Character Set Issues Relating to Computerization of The Eastern Han Chinese Lexicon《説文解字》*Shuowenjiezi* ". Hong Kong: Proceedings of the 18th International Unicode Conference.

DUAN Yucai 段玉裁 (1735-1815)
1815 《説文解字・注》。*Shuowen Jiezi - Zhu.* [SWJZ-Annotated; Qing recension.]〔東漢〕許慎著〔清〕段玉裁注。上海：上海古籍出版社，1989. ISBN 7-5325-0487-5/H.6.

ISO/IEC
2000 *Information technology - Universal Multiple-Octet Coded Character Set (UCS) - Part1: Architecture and Basic Multilingual Plane.* Reference Number ISO/IEC 10646-1 :2000(E). Second Edition, 2000-09-15. <http://www.iso.ch>.

ISO/IEC JTC1/SC2/WG2/IRG (Ideographic Rapporteur Group)
2000 *CJK Unified Ideographs Extension B*, for ISO/IEC FCD(R2) 10646-2:2000(E), Ordered by Kangxi Dictionary. 2000.12.05. <http://www.cse.cuhk.edu.hk/~irg/>.

JENKINS, John H. 井作恆
2001 "New Ideographs in Unicode 3.0 and Beyond". San Jose, California: Proceedings of the 18th International Unicode Conference.
1999 "New Ideographs in Unicode 3.0 and Beyond". San Jose, California: Proceedings of the 15th International Unicode Conference.

LU Qin 陸勤
2001 "The Ideographic Composition Scheme and Its Applications in Chinese Text Processing". Hong Kong: Proceedings of the 18th International Unicode Conference.

LUNDE, Ken 小林劍
1999 *CJKV Information Processing*. Beijing, Cambridge: O'Reilly. ISBN: 1-56592-224-7.
1999 "Acrobat 4.0 Adds CJKV Features: Embedding multibyte fonts in PDF files eases cross-platform use of documents". *Multilingual Computing and Technology*, Volume 10, Issue 6. <http://www.multilingual.com/>.

XU Kai 徐鍇 (920-974)
<974 《説文解字・繫傳》。*Shuowen Jiezi - Ji Zhuan.*〔東漢〕許慎著〔南唐〕徐鍇選。北京：中華書局，1987. ISBN 7-101-00060-6/H.7.

XU Shen 許慎 (58?-147?)
121 《説文解字》。*Shuowen Jiezi.*〔東漢〕許慎著。 [Lost: see XU Kai and XU Xuan.]

XU Xuan 徐鉉 (916-991)
c.987 《説文解字・附檢字》。*Shuowen Jiezi - Fu Jianzi.* [Common modern Hong Kong edition, with appended indices; in hardcover and paperback; 陳昌治 (1873).]〔東漢〕許慎著〔南唐、宋〕徐鉉校定。香港：中華書局，1989. ISBN 962-231-208-X.

Unicode Consortium 統一碼團
2000 *The Unicode Standard, Version 3.0*. Reading, Massachusetts: Addison-Wesley.

APPENDIX 3: TEXT-BASED TYPOLOGICAL ENCODING

Adequate historical typologization of the Chinese script can only be accomplished with reference to specific texts and inscriptions, and this is what I mean when I refer to a "Text-based" or "Source-based" encoding. These encodings may also be termed "Contextual" in that they seek to document the historical context whence the glyph usage derives. The written sources of ancient Chinese are many and varied, and each offers stylistic peculiarities and mapping challenges. Oracle Bone Inscriptions, Bronze Inscriptions, Stone and Earthenware Inscriptions ... all of these contain vital historical information which only a typological system can address.

Typologization adequate for Chinese purposes would be rather simple in some respects. First, it should characterize aspects of GLYPH SHAPE, and second, it should characterize aspects of GLYPH USAGE. One may however complicate and broaden this scheme in many ways, for example with script-specific issues of componential encoding. The following is an outline summarizing more generally applicable elements of some of the databases mentioned in my IUC-18 paper (q.v.)

- **Fields of TYPE 1, relating to GLYPH SHAPE:**

SHAPE:CLASS: Type [main,variant];
main [shape class(es),variants:list];
variant [main class,variant class];
valence [isolate,combining:list(compounds)].
element [elemental,compound:list(components)];
SHAPE:VALUE: Type [outline,bitmap];
outline [lines,curves,metrics,hinting,kerning];
bitmap [size,points,color];
render [size,position,rotation,orientation,style];

- **Fields of TYPE 2, relating to GLYPH USAGE:**

USAGE:CLASS: Type [print (or inscription),image,encoded];
print [bibliographic citation];
image [jpg,gif...;link];
encoding [codepoint(s)];
location [page.line];
stats [instance.frequency];
status in source [active,defunct...].
USAGE:VALUE: Type [orthographic,phonologic,phonetic,morphologic,semantic,syntactic...]
analysis [source data:description,variant class...].

APPENDIX 4: ACKNOWLEDGMENTS

The writing of this paper was supported in part by grants from:

- The National Science Foundation (NSF), Division of Behavioral & Cognitive Sciences, Linguistics, Grant No. BCS-9904950;
- The National Endowment for the Humanities (NEH), Preservation and Access, Grant No. PA-23353-99.

Thanks to John Jenkins and Ken Whistler for being tireless sources of the highest quality information on the developing *Unicode Standard*. Without their help I would know little of Unicode and understand even less (with their help, I take full responsibility for what I know of it and misunderstand). Thanks also to Tom Bishop of <http://www.wenlin.com/>. Tom's work on *Wenlin* is now as always a continued inspiration.