Unicode Chinese paleography: making the evolutionary leap from bone, bronze, silk, and paper, to electronic bits

Dr. Richard S. Cook
rscook@unicode.org
STEDT Project, Linguistics Department
University of California, Berkeley

Text of a talk presented in the
Early China Workshop Series
Department of East Asian
Languages and Civilizations
University of Chicago
February 8, 2005

CONTENTS

ABSTRACT

As more and more rare characters are encoded, Unicode provides better and better support for Chinese. In conjunction with CDL technology, texts of many historical periods can now be digitized with unparalleled accuracy. For even greater accuracy, a move beyond the encoding of modern-style CJK characters is required, and specialists from all over the world have begun to express interest in working out standards for digitization of paleographic Chinese materials of many periods. This paper seeks to provide information on methods for establishing standard frameworks for Chinese paleographic research, and to raise awareness of developing international standardization efforts.

In the slide-show presentation, I will discuss seven steps in digitizing paleographic materials, and introduce some of the computerized systems and tools created for and used in my 說文 *Shuo Wen* work, including the 文林 *Wenlin* CDL database. Key features of my expansive Unicode 4.X character variant mapping table will be shown, and the manner in which this table is key to situating my SW texts in various contexts is explained. Historical background on the SW textual tradition is given, and some aspects of the challenges involved in establishing intertextual mappings in a standard framework will be illustrated.

INTRODUCTION

Computer technology promises to revolutionize the sciences, and we have already begun to see this in the present day, in many disciplines. But before it can fulfill its promise, much standardization work remains to be done, and it cannot be done quickly.

Standards are essential to all sciences, but especially to computer science, which would serve as the basis for research and information exchange among all the other sciences. Computers provide us with fabulous tools for research and publication, but these tools will hobble us and our findings if they are not standardized. Information exchange is at the heart of science, and unless we put in place today the tools to ensure the widest circulation and long-term preservation of computer data, there is no hope that future generations will have access to the work we do today.

The Unicode Standard is one such tool, and XML is another. With Unicode and XML it is possible to encode and present your data in a way that will ensure its long-term survival and world-wide accessibility.

Chinese paleography is concerned with realms far removed from the modern day. Paleographers work on inscriptional and other materials from days gone by, attempting to bring the knowledge and conventions of those past times into the present day. And it is the fact that there were in those past days differing conventions, different standards, that makes this task so difficult and complex. Even if it is possible to link an old character in a specific old period to a modern character, in some satisfying way, chances are that there may be other possible ways to do the linking. No matter how the relations between old and new are established, the relation itself is often a matter of educated opinion rather than of incontrovertible fact, and opinions may vary.

Relations then, from old to new and from new to old, are contextually dependent, and there are multiple contexts. One context might be the specific edition of the old text that one is interested in computerizing. Another context might be the particular interpretive school that one is following in reading that text. Textual editions can vary, as can interpretations. A digital system for paleography must be able to handle variability in both of these dimensions, seamlessly and accurately, or information will be lost when it is exchanged, and the research findings will be vitiated as a result. A digital system for paleography must provide the framework for establishing meaningful relations among different textual editions, and provide a way to relate various interpretations of the underlying data. Where the data is “the same underlying data”, the system must provide a way to indicate this relation, without imposing changes on the underlying data itself.

These problems are multi-dimensional, and very complex in each of the multiple dimensions. To situate some of my introductory remarks in substantive context, I should like to focus on digitization issues relating to a specific text, and a specific edition of that text. With concrete examples drawn from this work, it is my hope that some of the larger issues can be made plain.

THE EASTERN HAN CHINESE GRAMMATICON

Just as establishing standards for information exchange is important in the present day, so too is it important in every period. When standards change over time, there is a disconnect between the past and the present, and information is lost. Perhaps there is always a disconnect. Change is natural in all natural systems: they develop over time, they degrade and they improve. Is it possible to build standards today that are dynamic in the face of such natural processes? Indeed, the standards which we would like to build today will be adequate if they are able to survive unchanged into the future. But is this really possible? Do we really want to stop all change dead in its tracks, or just some kinds of change? Certainly, one should not expect to eliminate all change, but is the expectation that it is possible to eliminate catastrophic change misplaced?

The underlying media through which we communicate are essentially stable in the long term, insofar as media endure in time, and continue to be used in different times. Where media degrade or fall into disuse, they must be translated into more enduring media. This is an ongoing challenge, and shall be a challenge forever, I think.

In translating our data out of bone, shell, shard, and bronze, out of bamboo, silk and paper, and into electronic bits, we are making a great evolutionary leap. We are moving out of the material, physical world, and into the world of pure energy. Evolutionary leaps are always fraught with danger. Just as when the child learns to walk — his move from quadrupedalism to bipedalism is punctuated with falling down and with skinned knees — so too we who would step from our old material culture into the new electronic one can expect to suffer along the way. But the promise of the new realm is what drives our migration, and we expect the benefits to outweigh the losses.

Just as we are able to communicate with each other today by means of our linguistic conventions, so too in the Chinese Eastern *Han* Dynasty there were communicative conventions. Communication in the different media employ different conventions. In terms of print communication, the Eastern *Han* period saw the codification of the first great standard for Chinese writing, in 121 A.D., with the publication of 說文解字 *Shuo Wen Jie Zi*, the Eastern *Han* Grammaticon.

Shuo Wen (SW) stands as the first great attempt to address the wide variation in the Chinese written language. At the time of its publication Chinese had already been written for more than 1,000 years. The textual evidence which it seeks to address is drawn from the whole of the received literary tradition, and so although it is synchronic in that it is the product of a specific time, it is diachronic in terms of its subject matter. The two strands of old and new are inexorably intertwined, as the lexicographer looks in old writing and seeks to translate it into the present.

Shuo Wen stands at this crux also as the basis for modern Chinese paleographic study, and one finds, for example, that bronze epigraphy collections such as the encyclopedic 金文詁林 *Jin Wen Gu Lin* take *Shuo Wen* as their organizational backbone.

As the archaeologist will proceed with his excavation by carefully brushing away and examining and inventorying the surface layers, so too in paleographic study we cannot simply leap from the modern day directly into the past. Proceeding right to oracle bone inscriptional evidence without first moving carefully through the intervening evidence is as foolhardy as moving into a priceless archaeological site with dynamite and bulldozers. The paleographic equivalent of the archaeologists brush may be a writing brush, but today it is a virtual electronic brush such as one might find in a graphics program. With digital tools, it is possible to precisely document the texts to which we refer in our research. We can take photographs of the texts, and produce indexes of their contents, with a degree of precision that has never before been possible.

In my work on *Shuo Wen*, I have sought to digitize this text in stages, to provide researchers with tools of the finest quality and of the highest reliability. I have worked to photograph specific textual editions, to index them in exquisite detail, and to establish the internal and external relations among them. Only on the basis of such painstaking work is it possible to build the framework for responsible paleographic study.

My initial *Shuo Wen* work focused over ten years on the 說文解字注 *Shuo Wen Jie Zi – Zhu* of 段玉裁 *DUAN Yucai*, that preeminent 清 *Qing* Dynasty (18th century) assessment of the textual tradition. This edition, so often

referred to by lexicographers and paleographers, was the *de facto* choice, and needed to be dealt with first, in that the path which it blazes into the difficult text is innovative and the first scientifically responsible one. Duan's treatment of the text is essentially an internal reconstruction of the 東漢 Eastern *Han* original, buttressed with expansive and insightful commentary on the full classical tradition, such as it was known in his day.

Having produced a careful digitization of Duan's SW text and commentary, I then began the difficult task of relating Duan's work to other editions of SW, and to other lexical works. A complete mapping table of 丁福保 *DING Fubao*'s encyclopedia of SW editions 說文詁林 *Shuo Wen Gu Lin*, was produced, as was a font and mapping table for one of the principal *Qing* editions of the received 徐鉉 *XU Xuan* (*Song* Dynasty) text. In addition, mapping data was refined for the 56,000 entries of 漢語大字典 *Hanyu Da Zidian*, one of the principal sources in the Unicode/ISO encoding process. Also, electronic mapping and MC phonological data was created for a specific edition of the 宋本廣韻 *Song Ben Guang Yun* text (including 反切 *fanqie* and IPA transcriptions), and for Karlgren's *GSR*. (Much of this data is currently available online, through the Unicode Consortium's public "Unihan" database, which I help to maintain, in collaboration with John Jenkins of Apple Computer.)

The Seal fonts created for digitization of these texts were then mapped to Unicode 4.x, in a process that drew me more deeply into work with the Unicode Consortium and with the ISO/IEC 10646 international standards community.

The primary tool employed in this mapping work was *Wenlin*, and the CDL (Chinese Character Description Language) was expanded and refined in collaboration with Wenlin Institute during this period. A large descriptive database was created, covering with unmatched precision the entire Unicode BMP (Basic Multi-Lingual Plane, plane zero of Unicode's 17 planes), and the relevant portions of higher-level Unicode planes. Unicode encodes some 71,000 CJK (Chinese, Japanese, Korean and Vietnamese) characters, and the repertory continues to grow.

The CDL system is key to the encoding process, in that it provides a precise stroke-based model in a Unicode/XML framework for describing any modern-style Chinese character, and therefore a means to quantify differences among such script entities. CDL also provides a means of associating up to 4 billion CDL descriptions with any Unicode code point, and so CDL is the only database system adequate to addressing the character variation issue.

DIGITAL PALEOGRAPHY IN SEVEN STEPS

My SW and CDL work highlights the fact that prior to doing paleographic study, one must have an adequate framework for representation of the modern script. CDL is just such a framework, bonded to graphical images and traditional fonts, and situated in text-specific mapping tables.

CDL is not, however, adequate for representation of Seal-style and other old characters themselves, except through the level of interpretation which gives them representation by means of modern "equivalents". CDL, using a set of basic stroke types situated in a standard grid-space, could however be extended to accommodate other script styles. Whether or not other script styles are as suitable to this kind of treatment is an open question, dependent upon the extent to which they make use of a basic set of stroke types.

A CDL for Seal-style characters, for example, might be defined by making an inventory of the basic strokes and shapes em-

ployed in such characters, but the set of elements so defined would in no way be in a simple one-to-one relation with those of the current "Song-style" CDL. For one thing, Seal characters have a much simpler set of stroke types, but "stroke" itself must be defined rather differently, since some Seal strokes are curved in ways that Song strokes are not, and Song strokes are "bent" in ways that Seal strokes are not. In fact, Song strokes exhibit a high degree of refinement of the notion of "stroke", for indexing purposes. And this refinement is simply not present in Seal characters.

In my earlier SW study (2003) I present a table of single-stroke Seal characters, but did not at that time attempt to extend CDL to adequately describe Seal characters directly. CDL's support for Bezier curves does however permit this as a possibility for future work. Similarly, CDL could be extended for use in any script at all. For example, the roman text elements in *Wenlin*'s CDL database are built in this very way, although the graphical primitives are used for the most part as simple drawing elements rather than as graphemic elements.

Since development of and need for a CDL representation is reliant upon a high degree of abstraction having been attained in refinement of the script elements themselves, a great deal of work must be done to determine the basic elements that a CDL system would use for a given script. We may start with higher-level units, and then analyze these elements for recurrent patterns for reuse across descriptions. There is no necessary correlation between the elements so identified and the elements which a writer might be thinking of when writing, though in the best case we should hope that the two sets might be highly similar. With *Song*-style text, for example, the definition of "component" is highly dependent on the current state of the glyph inventory: addition of a new character may suddenly cause some assemblage of strokes to be identified as a component, in that two characters now share this assemblage. And no single writer may ever have known both of these characters.

The seven steps in my digital paleographic method (*cf.* the "7 encoding levels" in my 2003 study, p. 334) are as follows:

1) An image of the text to be digitized must be created. Prior to this the following are required: an image format must be decided upon, an image capture tool chosen (camera or scanner, depending on the medium), and a method for archiving the images selected. You may need to consult with a graphics professional, for recommendations in this regard. With images on disk (and with multiple backups), one can then go to the next step.

2) All of the distinctive elements in that text must be extracted and inventoried in a database. Extraction of the glyph image from the base text, assigning it a unique name, and situating its ID in a database that can track its origin and other features is a major undertaking.

3) Fonts can then be created based on the extracted images, for use in text-processing applications. At this stage, if the script entities do not have a straightforward mapping to something already encoded in Unicode (and since I am talking about Chinese paleography here, I will assume they do not) a font developer will need to use Unicode's "Private Use Area" (PUA) code points. For example, my Seal characters are currently using some 25,000 code points in Unicode's top-most private-use plane.

4) If the character set defined in stage 3 is representative of a script of a particular period and of interest to a user-group, then we might consider at this stage preparing a proposal to define a standard encoding for it. This will permit us to no

longer use the impermanent non-standard (and therefore dreaded) PUA.

5) The Unicode mapping (PUA or non-PUA) prepared in steps 3 or 4 may however be used to create mappings to standard Unicode code points. For example, my Seal characters are one column in my Unicode 4.x mapping table, each PUA Seal form having one or more encoded "equivalents", according to various metrics for determining equivalency.

6) With the mapping tables and fonts in hand, we are now in the position to create in-line editable representations of our text. There are various levels of representation possible. The lowest level representation will employ the precise glyph extracted in the process of creating the initial comprehensive inventory. At higher levels, non-distinctive differences among glyphs are ignored, to create "unified" representations. At another level, we may use, for example, a standard mapping of some type, to represent the whole text.

7) By means of the standard mapping of a given type, it is then possible to establish the relations between this text and other texts.

Under step 4 is mentioned preparing a proposal to define a standard encoding. If one has developed an inventory of script elements that is based on generally accepted scholarship, and is well-defined both in terms of character set size and periodization, it might be appropriate to consider working to have it encoded in the Unicode Standard. Information about the Unicode script encoding model is available in the printed Unicode book, and on the Unicode web site, as is information relating to the encoding process itself.

Extension of the ISO/IEC 10646 CJK character repertoire is undertaken primarily by the Ideographic Rapporteur Group (IRG), comprised of delegates from businesses and organizations around the world. The most recent contribution from the IRG to ISO/IEC 10646 is the massive 40,000+ "Extension B" CJK character set added in Unicode 3.1, and Extension C1 is now in development.

This group has recently begun the process of refining the encoding model for CJK, so as to make use of stroke-based data such as that currently in CDL. It is Wenlin's hope to merge CDL into the international standards process, to create a public database resource, a tool for maintaining and improving access to the vast Unihan character repertoire. The IRG is also engaged in discussing the feasibility of encoding inscriptional and other paleographic forms derived from the long history of writing in China.

In general, proposals to add script support to Unicode should be prepared in accordance with the Unicode encoding model, following guidelines and using forms available from Unicode and ISO/IEC online. Because of the complexity of creating a unified CJK repertory, proposals to augment the CJK repertory are usually channeled through the IRG, though there have been recent exceptions involving synchronization of Unicode with other national standards.

CONCLUSION

The encoding of paleographic forms is even more complex than CJK unification, in part because the character sets to be encoded are not well defined and are not in common use. Writing in one historical period using one or more writing media blurs into another period using other writing media. For this reason, some have argued that standard encoding of these ancient character forms is not simply inappropriate, but entirely impossible to do in the manner of the current Unicode encoding model.

My own feeling is that either a new encoding model for archaic scripts is needed, or else the problem must be viewed from a completely different perspective. Glyph collections derivative of paleographic sources of all kinds should be created step by step, in a framework such as that which I have briefly described here. Through such a process we shall then have carefully documented our character inventories, linked them solidly to their real-world sources, and made them maximally useful and extensible in the long-term.

I would encourage people interested in blazing new trails in paleography to tread the digital path, and to get involved with ongoing methodological discussions. Contributing to and making use of developing standards is the best possible way to ensure the widest accessibility and long-term survival of your research findings.

REFERENCES

Cook, Richard S.

- *說文解字電子版 Shuo Wen Jie Zi - Dianzi Ban: Digital Recension of the Eastern Han Chinese Grammaticon*. Doctoral Dissertation, UC Berkeley: Dept. of Linguistics. May, 2003. UMI # 3105189. http://socrates.berkeley.edu/~rscook/pdf/Diss-abstract.pdf
- *The Etymology of Chinese 辰 Chén*. UC Berkeley: Monograph Volume 18.2 of LTBA, 1996. http://socrates.berkeley.edu/~rscook/html/LTBA-18.2.html
- Additional bibliographic materials are available here: http://socrates.berkeley.edu/~rscook/html/writing.html

The Sino-Tibetan Etymological Dictionary and Thesaurus Project (STEDT)
http://stedt.berkeley.edu/

The Unicode Standard 4.1.
http://www.unicode.org/

The Ideographic Rapporteur Group (IRG)
http://www.cse.cuhk.edu.hk/~irg/

Wenlin Institute's *Character Description Language* (CDL)
http://www.wenlin.com/cdl/

The research presented in this study was supported in part by the following STEDT Project grants:

- National Science Foundation, Division of Behavioral & Cognitive Sciences, Linguistics, Grant No. BCS-9904950
- National Endowment for the Humanities, Division of Research Programs, Grant No. PA-24168-02

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation or the National Endowment for the Humanities.