

Comments on the work of the Old Hanzi Group towards an encoding of OBI script.

Adam Smith (02/10/12)

Summary

0. Introduction

1. General remarks

There is a pressing need for a standardized encoding of Old Hanzi. Both Old Hanzi experts and the IRG can benefit from working together.

2. Other projects

The Old Hanzi Group needs to explain how its work relates to earlier or ongoing work by other groups with similar aims. *Yinxu jiagu keci leizuan* 殷墟甲骨刻辭類纂 (LZ) and *Jiaguwen zixing zongbiao* 甲骨文字形總表 (ZB) are the most important among these. Mappings need to be provided between the characters proposed by the Old Hanzi Group and those of previous projects. How does the work of the Old Hanzi Group relate to that of the recent “China Font” project (or “China Character Collection” project - ‘中華字庫’工程) sponsored by the General Administration of Press and Publication of the PRC (中華人民共和國新聞出版總署)?

3. Feasibility of OBI standardization and encoding

OBI encoding is feasible, given the current state of knowledge of the script. The success of existing projects, in particular ZB, show that this is the case. The Old Hanzi Group should aim at encoding a stable, well-established core of the script. Extensions are to be expected subsequently, to accommodate new archaeological discoveries and new research.

4. Why Old Hanzi should be encoded separately from CJK

Old Hanzi should be encoded separately from CJK. However, mappings between OBI and CJK should be provided.

5. Documentation

The work of the Old Hanzi Group is currently very inadequately documented. Amongst other things, this makes external review very difficult.

6. Intended user group, and purpose of encoding

The Old Hanzi Group has not clearly envisaged the needs of the user group it intends to serve. Doing so in detail would help it resolve many of the difficulties it currently faces (where, for example, to draw the distinction between a character with a separate encoding, and a graphic variant to be represented by some other means.) In particular, the Group needs to envisage the needs of those

engaged in encoding a text.

7. Sort order, indexing and *Shuowen Jiezi* radicals

The use of Shuowen radicals as an index has certain advantages. However, the Old Hanzi Group should consider indexing its character list by other means also. Most important would be mappings to ZB.

8. Characters, glyphs, variants, and principles for unification

The Old Hanzi Group appears to be aiming to encode impracticably subtle visual distinctions for OBI. The Old Hanzi Group needs to specify more clearly the definition of a **character** (to be granted an individual code point) and a **variant** (to be distinguished by some other means, if at all).

0. Introduction

This document is a response to an invitation by Suzuki Toshiya and Deborah Anderson to comment on work by the Old Hanzi Group towards an encoding of early Chinese scripts. The comments are based on a review of documents archived on the IRG website:

(<http://appsrv.cse.cuhk.edu.hk/~irg/>), and of the data deposited at <ftp://ftp.iso10646hk.net/IRG/OldHanzi/>.

My interest in the work of the Old Hanzi Group stems from a small ongoing project of my own to develop a system for the electronic transcription and analysis of early Chinese excavated texts. I am currently working on a database of transcriptions for the Huayuanzhuang Dongdi 花園莊東地 corpus of oracle-bone inscriptions.¹ This has entailed the development of an ad hoc encoding, a font, an input method and tools for working with the data.

Although I am broadly familiar with the aims and methods of the IRG, the Old Hanzi Group, and scholarship on early Chinese texts, the fact that I have not worked with the IRG or the Old Hanzi Group previously will inevitably mean that the comments offered here will reflect misunderstandings on my part of their activities, or be marred by a use of terminology that is not fully compatible with theirs. Nevertheless, I hope that the remarks will be useful. They reflect the perceptions of an informed outsider and likely future user of any encodings, meta-data, and tools developed through the work of the Old Hanzi Group.

1. General remarks

As pointed out in IRGN1014, “there is a pressing need for encoding Old (pre-Qin Dynasty) Hanzi” to support the scholarly transcription and analysis of texts, and their publication for both specialist and non-specialist audiences. I would add that it is especially desirable that the encoding be done with a rigorous attention to principles of unification, so that the full potential of computer text processing (indexing, searching, concordance building, input methods, etc.) can be realized. Since most scholars working on early Chinese texts have very little experience with these issues, there is an obvious role for the IRG to play. Transparent documentation, public review, and publication of meta-data used in developing the encoding are also highly desirable from an end-user’s perspective. Again, the IRG provides a framework to support those goals. For that reason, I hope that the Old Hanzi Group is able to sustain its working relationship with the IRG, and that the sometimes conflicting perspectives of philological scholarship on the one hand, and encoding standards on the other can fruitfully accommodate one another.

The work of the Old Hanzi Group has been focused on the script used in the late 2nd millennium BC divination records on bone and shell, the so-called “oracle-bone inscriptions (OBI)”. Except where noted, the scope of these comments is also limited to that body of material.

2. Other projects

Several other projects have already produced results that are immediately relevant to the goals of the Old Hanzi Group, or are likely to do so in the next few years. No mention is made of these other projects, their aims, or present results, in any of the Old Hanzi Group documentation that I have had time to review, with the exception of comments made by Japanese scholars who seem to be only on the periphery of the work being done by the Group.

Yinxu jiagu keci leizuan 殷墟甲骨刻辭類纂 (LZ) is a concordance to OBI texts, covering much but

¹ Zhongguo Shehuikexueyuan Kaogu Yanjiusuo 中國社會科學院考古研究所, *Yinxu Huayuanzhuang dong di jiagu* 殷墟花園莊東地甲骨 (Kunming: Yunnan Renmin Chubanshe 雲南人民出版社, 2003).

by no means all of the published corpus.² The character table (*zixingbiao* 字形表) which provides the index to LZ tacitly addresses many of the issues that an encoding must address: sorting via serial numbering; structural analysis of characters; unification or not of variants. The LZ character table has also been used to index other works, notably *Jiagu wenzi gulin* 甲骨文字詁林 (GL), an important compilation of analyses of characters in the OBI script.³ One of the members of the Old Hanzi Group, Zhao Cheng 趙誠 is the coauthor of LZ, under the pen-name Xiao Ding 肖丁.

The CHANT project has produced electronic transcriptions of much of the OBI corpus.⁴ These are accessible online, though only via a web interface that is too limited to perform analyses other than simple text searches. Many US research universities have subscriptions, and it is widely used by specialists. This is the only encoding of OBI about which this can currently be said. Two character tables have successively appeared as print publications associated with the CHANT project. *Xinbian jiaguwen zixing zongbiao* 新編甲骨文字形總表 is based on the LZ character table, and is in effect a revision of it.⁵ It includes mappings to the LZ serial numbering. It appears from casual inspection to be the character table used for the CHANT digital transcriptions as they currently appear online. *Jiaguwen zixing zongbiao* 甲骨文字形總表 (ZB) is a further revision and extension, incorporating additional characters that appeared for the first time in recently-published material.⁶ ZB is probably the most comprehensive and well-documented character table for OBI currently available. The CHANT project is supported by digital fonts capable of displaying OBI forms, and *kaiti* 楷體 equivalents that are not part of the standard CJK repertoire.

Other recent projects that have resulted in published character tables include work by Chen Tingzhu.⁷

It would be useful for any reviewer or end-user of the Old Hanzi Group's work-in-progress or final encoding of OBI script to understand how the Old Hanzi encoding relates or compares to the character tables produced by the other projects, especially ZB. Are they similarly comprehensive? Do they agree on the treatment of character variants? Can one be used to fill some of the gaps in another?

Although it would represent a considerable investment of time, I think that the Old Hanzi Group should publish a mapping between their character table and that of at least one other project's. ZB would be the best choice, since it has been through several revisions, it already includes mappings to LZ, and it has supported the large digital transcription project mentioned above. A statement by the Old Hanzi Group about how its character table will relate to, supersede or complement those of other projects would also be desirable.

The task of gathering character (glyph?) exemplars is a large and complex task. There are 9,572 rows in the current version of the Old Hanzi database (OldHanZi_20110214.xls). Presumably the Old Hanzi Group is basing their work on an existing character table. Details about this underlying table, and

2 Yao Xiaosui 姚孝遂 and Xiao Ding 肖丁, *Yinxu jiagu keci leizuan* 殷墟甲骨刻辭類纂, 3 vols. (Beijing: Zhonghua Shuju 中華書局, 1989).

3 Yu Xingwu 于省吾, *Jiagu wenzi gulin* 甲骨文字詁林, 4 vols. (Beijing: Zhonghua Shuju 中華書局, 1999).

4 “漢達文庫 CHANT (CHinese ANcient Texts) Database”, n.d., <http://www.chant.org/>; Che Wah Ho, “CHANT (CHinese ANcient Texts): a comprehensive database of all ancient Chinese texts up to 600 AD,” *Journal of Digital Information* 3, no. 2 (2002): no page #s, <http://journals.tdl.org/jodi/article/viewArticle/81/80>; Chen Fangzheng 陳方正, “讓甲骨文字走向大眾 (CHANT project)”, n.d., http://www.chant.org/news/issue5/fc_i.asp.

5 Shen Jianhua 沈建華 and Cao Jinyan 曹錦炎, *Xin bian jiaguwen zixing zongbiao* 新編甲骨文字形總表 (Hong Kong: Zhongwen daxue chubanshe 中文大學出版社, 2001).

6 Shen Jianhua 沈建華 and Cao Jinyan 曹錦炎, *Jiaguwen zixing zongbiao* 甲骨文字形總表, Revised edition. (Shanghai: Shanghai Cishu Chubanshe 上海辭書出版社, 2008).

7 Chen Tingzhu 陳婷珠, *Yin-Shang jiaguwen zixing xitong zai yanjiu* 殷商甲骨文字形系統再研究, Di 1 ban. (Shanghai: Shanghai Renmin Chubanshe 上海人民出版社, 2010), 469–614.

how it is being extended and revised, would help an external reviewer to understand the group's methodology.

In the summer of 2011, the General Administration of Press and Publication of the PRC (中華人民共和國新聞出版總署) announced the “China Font” project (or “China Character Collection” project - “中華字庫”工程).⁸ The stated aim of the project is to encode all Chinese characters (*hanzi*) and all characters in ethnic minority scripts, as well as to develop supporting text-processing technologies. The project expects to encode 500,000 characters, including 100,000 Old Hanzi (*hanzi guwenzi* 漢字古文字), 300,000 *kaishu* 楷書 forms, and 100,000 ethnic minority characters. The project is expected to be completed “within five years”, with the involvement of “nearly 30 institutions of higher education, research institutes and businesses.”

The “China Font” project is likely to have a big impact on the infrastructure for Old Hanzi scholarship and publication. It is not clear, though, how the Old Hanzi Group sees its relationship to the “China Font” project. Is the Old Hanzi Group going to adopt the results of the “China Font” project for OBI encoding? Or vice versa? Or are the efforts of the two groups already merged? Or do the two projects have distinct goals? I note that Li Guoying 李國英 is a member of the Old Hanzi Group, and also a participant in the “China Font” project.

3. Feasibility of OBI standardization and encoding

The Old Hanzi Group's documents refer occasionally to the incomplete state of scholarly knowledge of Old Hanzi (for example, when advocating the encoding of Old Hanzi without unifying with CJK). This might invite the objection that current understanding of the script is insufficiently mature for a standardized encoding to be feasible. I would suggest, however, that the work of other projects has already shown the feasibility of OBI standardization and encoding.

The OBI character tables produced in association with the CHANT project, ZB in particular, are very comprehensive in terms of their coverage of the OBI character repertoire. Although the documentation that accompanies them is largely silent on the subject, they appear to be inspired by a rational approach to the handling of “abstract characters” vs. glyphs, and to the related problem of variants. The character tables have been fundamentally stable since being inherited from LZ, while accommodating successive revisions and extensions for newly-published inscriptions. The ZB character tables are in themselves a compelling argument for the feasibility of OBI encoding.

The changing state of scholarly knowledge should mean that the encoding of OBI is tackled by first encoding what is believed to be the stable core of the OBI character repertoire, with the expectation that extensions be produced in the future. Even if knowledge of the current corpus were perfect, the publication of newly-excavated material is always likely to make extensions necessary. The recent discovery of the Huayuanzhuang Dongdi corpus, for example, introduced many new forms.⁹ Although some of these could be understood as “variants” and unified with previously-recognized OBI characters, others have so far resisted unification (fig. 1).

8 “‘Zhonghua ziku’ gongcheng qidong ‘中華字庫’ 工程啟動,” General Administration of Press and Publication of the People's Republic of China 中華人民共和國新聞出版總署, July 27, 2011, <http://www.gapp.gov.cn/cms/cms/website/zhrmghgxwcbzsww/layout3/header.jsp?channelId=1025&siteId=21&infoId=720733>.

9 Zhongguo Shehuikexueyuan Kaogu Yanjiusuo 中國社會科學院考古研究所, *Yinxu Huayuanzhuang dong di jiagu* 殷虛花園莊東地甲骨.

Fig. 1 - new non-unifiable character from Huayuanzhuang Dongdi (ZB p. 38)

0463



HD 003

If the Old Hanzi Group were to work with the assumption that future extensions to the encoding would certainly be necessary, this might also encourage postponing difficult classificatory decisions, particularly those concerning subtle or marginally-attested variations. My sense is that the work of the Group is being hindered by classificatory issues of this kind, and that postponing them would accelerate progress towards encoding a stable core of the OBI script.

4. Why Old Hanzi should be encoded separately from CJK

The Old Hanzi Group has advocated encoding OBI separately from CJK, that is, they do not advocate unifying OBI characters with their CJK equivalents, even when the identification of an equivalent is unproblematic. Doubts have been raised about this decision by Richard Cook. On this issue, I think that the Old Hanzi Group policy is the correct one.

Many OBI characters can be unproblematically mapped to CJK characters. Many others, probably the majority, cannot. Keeping meta-data on mappings from OBI to CJK is very important, and I don't think this has been adequately achieved by the Old Hanzi Group. Nevertheless, unification with CJK is undesirable for the following reasons:

1. Continuity in character identities between OBI and CJK is complicated by splits and mergers, to the extent that identity with a single CJK form is problematic even when the identity of the character within the OBI repertoire is unproblematic.
2. There are multiple principles according to which an OBI character might be mapped to a CJK equivalent – linguistic value, nearest functional equivalent, exact structural match, etc. Advocates of unification with CJK are presumably thinking in terms of the latter, a structure-preserving continuous morphing of the character shape. If unification were the goal, this would indeed be the way to go. However, equivalences of this kind are very rarely used in scholarly practice – mappings to less exact structural equivalents, or to nearest functional equivalents, or to characters with the same linguistic value are much more familiar to practitioners. I foresee big challenges in conveying the importance of the strict structure-preserving mapping, required for meaningful unification, to members of the Old Hanzi Group, or any other group of conventionally-educated scholars working in the field.
3. Unification would require either a/ the creation of numerous “Song-style” glyph exemplars for those OBI characters that do not have exact CJK equivalents, or b/ a character table for OBI that contained a mixture of Song-style (unified with CJK) and OBI-style (no CJK equivalent) glyphs. Neither option seems particularly appealing. Moreover, trying to make strict structure-preserving Song-style exemplars would, I suspect, lead to glyphs that were unrecognisable either as OBI or as Song-style characters.

To illustrate, consider the high-frequency OBI graph 𠄎. Most OBI scholars asked to pick a CJK equivalent would offer either 貞 (which writes the same linguistic value – zhēn “to ask through divination” - in received literature) or 鼎 (a pictogram for dǐng “tripod or quadripod cooking vessel”, the source of the graph). Neither is an exact structure-preserving equivalent. The latter is indeed the ultimate source for the graph, but in the OBI script, 鼎 and the character in question have already undergone a split and need to be encoded separately. The latter has already added 卜 as a component, and substituted 貝 for 鼎 (a substitution that affects the history of other graphs also, including 則). Ideally, meta-data would allow the OBI graph to be retrieved by an appropriate search for either 鼎 or 貞. But unification with either CJK character would be inappropriate. It is not obvious to me what a “Song-style equivalent” of this graph would like, given that 鼎, 貞, and 貝 are all out of the question.

Meta-data for the character in question should probably include best linguistic match to CJK (貞), and graphic best match to CJK (鼎?). Certainly, end users might want to be able to search for this character using either of these CJK forms.

5. Documentation

I have an overwhelming sense that the work of the Old Hanzi Group is not being adequately documented. Definitions are obscure and often contradict one another. Revisions get proposed that would fundamentally reverse an existing definition, yet no clear statement is made as to whether the revision was adopted or not, and why. The methodology for sourcing the glyph exemplars that appear in the current spreadsheet is not at all clear. What steps does the Group take, for example, to ensure that an item that appears in the ZB character table also has an exemplar in their list? Given that we are talking about thousands of items in each table, and two different sort orders, this seems like a problem that would need to be carefully addressed. I’ve no reason to think that the Group has overlooked questions of this kind, but the documentation should reassure a reviewer that this is the case.

Perhaps the most worrying gap in the documentation is the absence of any statement about the envisaged format of the meta-data that would be necessary to support the encoding. The Group is pursuing an encoding that aims to represent very slight distinctions between forms. Any end-user attempting to enter a text in an encoding of this sort would need to have a sophisticated input method supported by meta-data. Such meta-data would probably need to include tables of variants, character component analyses, character frequencies, linguistic values, mappings to CJK and to published character tables, character co-occurrence patterns, and so on, in order for the input method to be viable. To embark on an ambitious encoding project without a clear plan for the collection and processing of meta-data strikes me as insecure.

The need to draft documents in English is evidently also a problem. Perhaps the Group could recruit additional assistance to ensure that the more important documentation is better drafted. The shortcomings in documentation are not principally linguistic, but nevertheless there are crucial places where the language is barely comprehensible.

6. Intended user group, and purpose of encoding

The intended user group and the purpose of the encoding have not been clearly identified. A request for clarification as to the intended user community (IRGN1522) elicited the reply “all in the world who are interested in learning and studying more about Old Hanzi.” (IRGN1524) But the question was not about who the ultimate beneficiaries of the encoding might be. One hopes that most of those interested in learning about Old Hanzi will never have to worry about how it is encoded. Rather, the question (as I understand it) is about the very much smaller user group that will actually have to manipulate the encoding, and about what they will be trying to do with it.

As I suggested in the previous section, the Old Hanzi Group has not clearly envisaged the needs

of the scholarly end-user who will need to input and search OBI text, and be able (one hopes) to perform more sophisticated corpus analyses. The very subtle distinctions that the Group proposes to encode will, most of the time, be of no interest to this user: if one is studying syntax, whether a stroke appears as a dot or a line, or whether a character appears in mirror image, is unimportant. Either scholarly end-users will need to develop applications to abstract the information of interest to them from the encoding, or else others with a good understanding of scholarly needs will need to develop and maintain these applications on their behalf. Those are the kinds of user roles that need to be more explicitly envisaged and accommodated when making decisions about which distinctions to encode or which meta-data to collect.

7. Sort order, indexing and *Shuowen Jiezi* radicals

The Old Hanzi Group has consistently advocated the use of *Shuowen* radicals to index the OBI character table. This choice has been questioned by reviewers (IRGN1346), who have proposed the radical system of LZ as an alternative. There is a great deal to be said for both alternatives, and the obvious solution is to include both as fields in the database.





Indexing by *Shuowen* radical facilitates not just referencing of the Han lexical text itself, but also referencing of numerous other Old Hanzi character tables that are already available, covering not just OBI but other text corpora as well.¹⁰ Having a *Shuowen* radical index will thus facilitate developing tables of correspondences between OBI and later script stages. The LZ radicals lack this useful property.

As pointed out in IRGN1346, using *Shuowen* radicals as the only index has considerable shortcomings. Many OBI characters that do not appear in the *Shuowen* can only be assigned a *Shuowen* radical arbitrarily, and many cannot be meaningfully assigned one at all. The coverage of the LZ radical index is better in this respect.

The procedure used by the Old Hanzi Group for assigning *Shuowen* radicals to OBI characters sometimes leads to counterintuitive results. For example, the OBI character in fig. 2, which arguably could be equated to any of 史, 事 or 吏, is assigned the radical — on the basis of the *Shuowen* radical for 吏, despite the fact that — is never a component of the OBI character.

10 E.g. Zhongguo Shehuikexueyuan Kaogu Yanjiusuo 中國社會科學院考古研究所, ed., *Jiaguwen bian* 甲骨文編, Kaoguxue zhuankan 考古學傳刊 Yi zhong di shisi hao 乙種第十四號 (Beijing: Zhonghua Shuju 中華書局, 1965); Rong Geng 容庚, *Jinwen bian* 金文編, Yingyin ben 影印本. (Beijing: Zhonghua Shuju 中華書局, 1985); Gugong Bowuyuan 故宮博物院 and Luo Fuyi 羅福頤, *Gu xi wen bian* 古璽文編 (Beijing: Wenwu Chubanshe 文物出版社, 1981); Teng Rensheng 滕壬生, *Chu xi jianbo wenzi bian* 楚系簡帛文字編 (Wuhan: Hubei Jiaoyu Chubanshe 湖北教育出版社, 1995).

Fig. 2 – a counterintuitive *Shuowen* radical (IRG N1747).

Imitation Script/Glyph	Original Script/Glyph	S.W. Radical
		—
		—

Similarly, the character in fig. 3 has been assigned the radical 邑 (IRG N1747), despite the fact that 邑 is not a component. As a solution to the problem of indexing this seems rather clumsy.


Fig. 3 – 邦 variant.



If *Shuowen* radical assignments like the two just mentioned are what is required to bring the OBI character table into alignment with existing *Shuowen*-indexed tables, then they are justified. Nevertheless, additional indexes would be highly desirable. An index based on the numbering used in the ZB character table strikes me as an absolute priority, as I have already suggested.

These so-called “radical” indexes involve the selection of one component of a compound character, and that selection is often to a degree arbitrary, and often difficult to predict for someone unfamiliar with the indexing system. I would also like to be able to search an OBI character table on the basis of all character components, not just those that have been deemed “radicals”. However, since this would involve a many-to-many mapping between characters and components, it would require some form of relational database instead of the flat file that is currently being used to present the character table.

As an illustration of the shortcomings of relying on a character table indexed by *Shuowen* radical, let me offer the following three examples.

Example 1 - 

This is the character that appeared in fig. 1. It appears in the Huayuanzhuang Dongdi corpus, first

fully published with an accompanying character table in 2003, which is supposed to be within the scope of the Old Hanzi group project (corpus C in IRG N1747). The character appears as number 0463 in the ZB character table, classified under the WOMAN radical 女. It occurs 22 times in my own transcriptions of the Huayuanzhuang corpus, so it is certainly well-attested. We don't know its phonological value and have to guess at its semantics, but that is true of many OBI characters. Its graphic structure presents no difficulties. It should be in the current Old Hanzi Group list (OldHanZi_20110214).

But which “radical” should I look under? I tried 女 – no luck. 林? 木? 虎? 虍? 夔? It is not listed under any of these radicals. After a burdensome search we have failed to find the character, but without feeling fully confident that the character is not in fact in the list somewhere, under some more obscure radical (remember that a radical does not actually have to be a component!). Since in this case we happen to know that the character has appeared in only a single published corpus (corpus C), we can search through all the table rows that have a value for the “source” column that begins “C-”. Again, nothing. We can probably conclude that the character is missing from the list. Had a mapping to ZB been provided, we would have had a much easier time determining that the character was absent, and its omission would probably already have been noticed by the Old Hanzi Group.

Example 2 – ORS04125-04128 vs. ORS07695-07696 (Fig. 4)

The character in question is (approximately) a compound of the SHACKLES pictogram 𠂔, and TIGER-HEAD/MAN 虍. Should the radical be the former or the latter? The Old Hanzi Group list has it under both radicals – an error according to the conventions they have established, but an easy one to make since the classification involves an arbitrary decision. Note that exemplars ORS04128 and ORS07695 are the same character in the same inscription, though the thumbnails come from different electronic images. Clearly one of them needs to be deleted from the list.

Fig. 4 – which radical?

ORS07 695	G03849			A-26972	商代	河南安阳	甲骨	𠂔	397	執	執		脚上铐上木铐	Keep
ORS07 696	G03850			A-26982	商代	河南安阳	甲骨	𠂔	397	執	執		是虎頭，被铐著的虎	Keep
ORS04 125	T02022			A-27302	商	河南安陽	甲骨	虍	168	虍	虍 ₁			
ORS04 126	T02023			A-26979	商	河南安陽	甲骨	虍	168	虍	虍 ₂			
ORS04 127	T02024			A-27281	商	河南安陽	甲骨	虍	168	虍	虍 ₃			
ORS04 128	T02025			A-26972	商	河南安陽	甲骨	虍	168	虍	虍 ₄			

The error could have been avoided by indexing characters under all identifiable components, using a many-to-many relation. The list of exemplars for the character in question would then be identified as precisely those with only the 2 components 𠂇 and 𠂆. There would be no danger of exemplars being filed under the “wrong” radical. A mapping to the ZB character table would probably also have prevented this problem.

The fact that this second example was encountered by accident, while researching the first example, and not through a systematic search for errors, makes one suspect that there are many more problems of this kind in the database.

Example 3 – ORS00230-ORS00231 (fig. 5)

Fig. 5

ORS00230	T00154			A-27221	商	河南安陽	甲骨文	示	003		蒸		[T_Oct]:OK[T_Oct]:OK . .	Keep
ORS00231	T00142			A-27632	商	河南安陽	甲骨文	示	003		蒸		[T_Oct]:OK .	Keep

The second of the two exemplars in fig. 5, ORS00231, has been assigned the radical 示, since 示 is a component. However, there is no corresponding *Shuowen* graph. The first exemplar, ORS00230, has also been assigned the radical 示 because the database editors believe that it is writing the same linguistic value (蒸) as the second exemplar. They may be correct about that, but indexing the database in this way means that a user would be unable to locate the first entry using the radical index unless they knew that there was an alternative writing of the same linguistic value that was indexed under 示.

Presumably these two exemplars have been placed in adjacent positions in the database because they are “structural variants” capable of writing the same linguistic value. If that is the case, ORS00198-9 and ORS00236-9 should also be moved to positions adjacent to ORS00230-1.

8. Characters, glyphs, variants, and principles for unification

In several documents, the Old Hanzi Group has laid down rules about how to encode variation in character forms. How different can two forms be and still be assigned to the same code point? What kinds of variation between forms should lead to their being encoded separately?

In addition to this distinction corresponding to the level of the assigned code point, there also seems to be some interest in documenting subtler variation in character forms, which will not be reflected in the code-point assignments, but which may nevertheless be a part of the encoding standard, or part of its supporting documentation or meta-data, or at the very least part of the process of developing the standard. In IRG N1522 Japan specifically raised the question of whether to employ compatibility characters or Ideographic Variation Selectors¹¹ to deal with this finer variation.

For the remainder of these comments, I will refer to written forms that will correspond to different code-points (in the future standard) as **different characters**. Written forms that are to be represented by the same code-point but which nevertheless have a recognized difference within the standard or its documentation I will refer to as **variants of the same character**. I will refer to an actual instance of a character as an **exemplar**.

The task of the Old Hanzi group is to come up with a set of rules that will allow end users to

¹¹<http://www.unicode.org/reports/tr37/>.





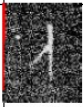
















identify any (legible, previously attested, etc.) **exemplar** as an instance of exactly one encoded **character** in a predictable way. What their task is with respect to variants depends in part on their answer to Japan’s question in IRGN1522 (2.3), which has not yet been satisfactorily addressed.

The Old Hanzi Group has collected a large number of **exemplars** in the form of thumbnails appearing in the “Original Shape/Glyph” column of the database. Evidently, there is still some discussion ongoing about which rows of the database should be retained and which deleted.

From IRG N1771R, it would appear that Japan thinks of each row in the database as typically representing a candidate **variant** (a “glyph”), rather than a candidate **character**. Most of the documents produced by the Old Hanzi Group, however, clearly imply that they intend the opposite: unless the “State” column is “Unified” or “Deleted”, every row is a **character**, to be given its own code-point. That would seem to mean that they do not envisage including **variants** (in my sense) within the standard.

If I am correct about the intentions of the Old Hanzi Group in this regard, then either the database needs some drastic revisions or the project to develop an encoding standard for OBI is in serious trouble. No conceivable purpose could be served by encoding 卜 bū (the DIVINATION CRACK pictogram) as eight different **characters** (fig. 5). If the notion of “abstract character” means anything at all, all of these exemplars should be encoded as a single **character**.

Fig. 5 – Seven of the eight database exemplars for 卜 bū.

ORS02 501	T01226			A-01027-0	商	河南安陽	甲骨	卜	094		卜			Keep
ORS02 502	T01227			A-33542	商	河南安陽	甲骨	卜	094		卜			Keep
ORS02 503	T01228			A-17673	商	河南安陽	甲骨	卜	094		卜			Keep
ORS02 504	T01229			A-33429	商	河南安陽	甲骨	卜	094		卜			Keep
ORS02 505	T01230			E-H11-38	周	周原	甲骨	卜	094		卜			Keep
ORS02 506	G01209			A-00006	商代	河南安陽	甲骨	卜	094		卜			Keep
ORS02 507	G01210			A-00005	商代	河南安陽	甲骨	卜	094		卜			Keep

It would also be difficult to make a strong case for including distinct **variants**, I would say. Whether the side arm of the 卜 pictogram points left or right depends on the orientation of the crack on the bone or plastron on which the inscriptions appears (and becomes standardized to point right in later texts). The angle of the side arm can take a continuum of values, and is subject to no obvious constraint (fig. 6). The burden of trying to transcribe a text with the eight distinctions that the database seems to be advocating would be heavy, and the result unreliable and of little value

Fig. 6 – exemplars of 卜 from the Huayuanzhuang Dongdi corpus.

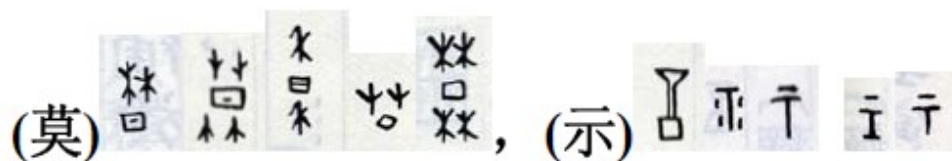


The example of 卜 is admittedly an extreme one, but even the exemplars shown in fig. 4 above should certainly be unified as a single **character**. There is clearly some variation in the way the exemplars are written, which could perhaps be captured in the proposed standard as **variants** (though I can envisage no purpose for doing so). Some of the TIGER-HEAD/MAN 虎 components have two arms; some have one; in some cases the arms extend into the SHACKLES 牵; in others, not. If the Old Hanzi Group really intends to encode this variation with more than one **character**, they need to imagine an input method that confronts the user with with all of these options in a clear and efficient way, and they need to provide the meta-data that a developer would need to implement it (i.e. verbal statements of what the distinguishing features of each exemplar are).

The Old Hanzi Group documents use the terms “variant constructs,” “construct variance,” “construct-variances” (IRG N1087 and subsequently), and “Yi-Gou [異構]” (IRGN1760 p. 19) to mean differences that are big enough to justify encoding as two separate **characters**. “Construct-variances are deemed as different ancient characters, thus they should be encoded separately.” (IRG N1087) Early documentation did not define these terms precisely, but the implication was that very subtle differences, of the kind discussed in the previous paragraphs, would be sufficient to entail “construct variance” and therefore separate encoding.

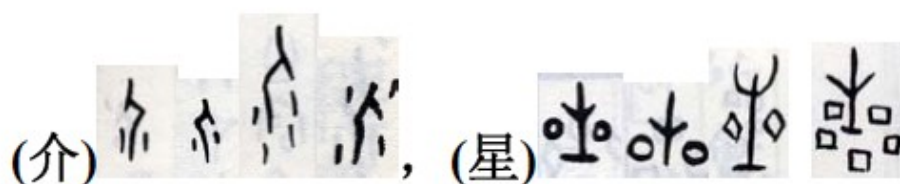
More precise criteria are provided in IRGN1271 (section 4), reaffirmed in N1524. However, by then the terminology had changed, and it is no longer made explicit that these are criteria for separate **characters**, as opposed to separate **variants**. The document simply states that these are criteria for deciding when “two or more instances of Old Hanzi are considered different from each other”. The implication, however, is that these are to be encoded separately. Let’s briefly look at the specified criteria:

4.1.1 One or more types of components are different.



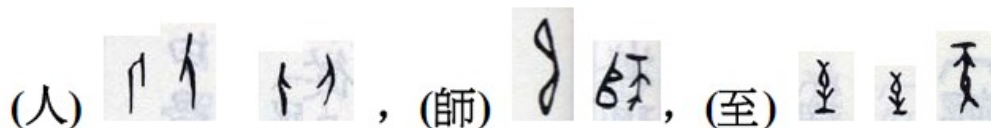
4.1.1. Apart from the 3rd and 5th forms for 示, which seem indistinguishable, the variation involved is objectively specifiable. In some cases, it is also structural, in the sense that the list of sub-components of the graph is different. The 1st and 5th 莫 have 木 as a component, while the others don't. Distinguishing some of these variants might be useful for certain sorts of analysis, but I am not fully persuaded that they should be encoded separately. Ideographic variation selectors may well be a better solution.

4.1.2 The number of components or lines is different.



4.1.2. The first two exemplars seem to be merely mirror images of one another (see below). Again, I am not persuaded that the considerable burden incurred by encoding these variants separately can be justified. Anyone entering OBI text would need to be aware that there are (at least) four variant writings for 星, and would need to be able to determine unambiguously into which class to place any exemplar he or she encountered.

4.1.3 The direction (e.g. mirror image) of a component is different.



The proposal of criterion 4.1.3 is simply a mistake. Most left-to-right mirror-image contrasts are determined by changes of text direction (boustrophedon). Most left-right asymmetric characters are subject to this in the early inscriptions. Surely no one wants to have two code-points for each of these.

4.1.4 The position of one or more components is different.



4.1.4. Once mirror-image variation (see above) is factored out, I can't see any objectively specifiable variation here. This should be encoded as a single **character**. I can't even imagine that it would be useful to define **variants** for these five forms either.

4.1.5 Whether the same set of components are connected each other or not.



4.1.5. Again, I'm not persuaded that the distinction is useful enough to justify the burden of encoding the two forms separately.

4.1.6 One or more line types (straight line, curve, circle, rectangle, closed line or curve filled inside) are different.



4.1.6. Having some variant glyphs available to anyone setting a text for publication might be useful, but having (at least) six different encodings for OBI 王 would simply be a nightmare. The problem is not just the burden of handling these during data entry or text processing, but also that the boundaries between the six (or more) categories would have to be objectively and unambiguously specified. I'm not persuaded that this is possible.

In summary, the Old Hanzi Group is advocating encoding subtle visual distinctions that are a/ impractical for data entry and text processing, b/ not objectively specifiable, and c/ not compellingly useful for scholarly purposes (purposes which have not, in any case, been explicitly described by the Group). Subtle visual distinctions, such as whether a region of an exemplar is filled in or not, can be retrieved by consulting an image of the inscription.

I suggest that the Old Hanzi Group's work would be improved by engaging more positively with the suggestions put forward in e.g. N1771. This does not mean that those proposals are perfect, but they

do provide candidate answers to questions that the Old Hanzi Group needs to address.