Universal Multiple-Octet Coded Character Set UCS

ISO/IEC JTC1/SC2/WG2 IRG N1747

Date: 2011-04-14

Title:	**Old Hanzi Principles and References (Version 3)**
Source:	**Old Hanzi Experts Group**
Status:	**Input to IRG**
Action:	
Distribution:	**IRG Members and Ideographic Experts**
Reference:	
No. of pages:	
Medium:	**Electronic**

The Oracle Bone principles and references (version 3) extracted and compiled from the following documents:
IRG N1135R, IRG N1182, IRGN1215, IRG N1267, IRG N1271 (version 2), IRG N1325, IRG N1460, IRG N1747A.

1. Format of submission

ID	Imitated Script/Glyph	Original Shape/ Glyph	Source	Period/ Epoch	Area /Site	Material	SW Radical	SW Radical Number	* Corresp. Modern Char (UCS Code)	* Unifiable Shapes	* Notes
1											
2											

2. Definitions

2.1. ID: It is the unique id that consists of one or two letters member id (G, T, K, KP, J, V, S, H, M) followed by four digit sequential numbers assigned by submitters.

Example: T0001 is one IRG global unique ID assigned to an Oracle Bone submitted by TCA.

2.2. Imitated Script/Glyph: The truthful trace from ‘Original Shape/Glyph’
[摹寫字的定義：忠實摹寫原形的字形。]

2.3. Original glyph: The glyph selected according to the principles of Oracle Bone selection (in the item 4) in this document.
[原形的定義:根據選字原則(本文件第 4 章) 从原拓選定的字形。]

2.4. SW Radical: The picture of ShuoWen Radical in Kai-style. The submitter is not required to provide the picture, it is assigned by the project editor.

2.5. SW Radical number : 1 – 540. The order is defined by 漢•許慎 ‘說文解字’ (大徐本).

3. Rules

3.1. The ‘Imitated Script/Glyph’ should be of standardized size and dimension.

3.2. The ‘Imitated Script/Glyph’ shall be in EPS format (resolution 1024×1024).

3.3. The format of bitmap images for imitated glyphs, original shapes and 540 SW Radicals are specified as follows:
128x128，Black and white bitmap

3.4. The last three columns are ‘optional’ field and they are indicated with an asterisk “*”, the other fields are mandatory.

3.5. The field “Corresp. Modern Char. (UCS Code)”

3.5.1. The field “Corresp. Modern Char. (UCS Code)” is filled by single codepoint or a list of CJK Unified Ideographs separated by semicolon (;).

3.5.2. Only URO (CJK Unified Ideographs in BMP without Extension A and CJK Compatibility Ideograph) characters can be used for the convenience of sorting or finding a character from the database. If there is no corresponding modern character, or exists but not coded in URO, the field must be blank. The note field should be used for the description for such cases.

3.6. If the field “Corresp. Modern Char.(UCS Code)” is blank, then the “Notes” field must be filled with justifications to indicate the glyph is well-understood, for example, the meaning of the ‘Imitated Script / Glyph’. Also “Note” field can include the description of the glyph structure when the glyphic components have the corresponding modern character in UCS.

3.7. Source: The “Source” field is the important key to exclude exactly duplicated data.

The “Source” field consists of two data elements with one optional element. They will be concatenated with hyphen character ‘-’.

- (mandatory) The 1st letter indicates the book reference number. The possible values are:

  (A) stands for《甲骨文合集》郭沫若主編, 中華書局, ISBN 9787101016536 (13 volumes), 1978-1982

(B) stands for《甲骨文合集補編》彭邦炯主編, 語文出版社, 1999, ISBN 7801264967 (7 volumes), 1999

(C) stands for《殷墟花園庄東地甲骨》中国社会科学院考古研究所, 雲南人民出版社, ISBN 9787222038776, 2003

(D) stands for《山東濟南大辛莊甲骨》**TBD**

(E) stands for《周原甲骨文》, 曹瑋, 世界圖書出版公司北京公司, 2002, ISBN 9787506256650

(F) stands for《小屯南地甲骨》考古学専刊乙種 18 号, 中国社会科学研究院考古研究所, 中華書局, 上下冊 (1980 and 1983)

(G) stands for《英國所藏甲骨集》李学勤, 中華書局, ISBN 9787101009569, 1992

(H) stands for《懷特氏等所藏甲骨文集》**TBD**

(I) stands for《ひとものこころ 天理大学附属天理参考館蔵品》Vol. 1, Num. 5, 道友社, ISBN 4-8073-0254-X, 1986

(J) stands for《德瑞荷比所藏一些甲骨錄》"Several Collections of Oracular Inscriptions in Germany, Switzerland, The Netherland, Belgium" (Fr. Jean Lefeuvre), Ricci, ISBN 9782950560247, 1997

(K) stands for《瑞典斯德哥爾摩遠東古物博物館藏甲骨文字》李学勤, 中華書局, ISBN 7101022561, 1999

- (mandatory) Oracle Bone number (甲骨拓片的編號) which consists of 5 digits assigned uniquely to each Oracle Bone.

- (mandatory) Oracle Bone number (甲骨拓片的編號) which consists of 5 digits assigned uniquely to each Oracle Bone.

- (optional) Identifier to determine the side of the Oracle Bone which consists of 1 digit. The possible values are '0' for front side, '1' for back side. If an inscription is carved only on one side, this element will be omitted.

Three examples of the "Source" field are listed below.

- A-00001 (does not have front and back side)

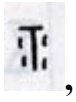
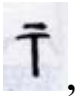
- A-00001-0 (front side)

- A-00001-1 (back side)

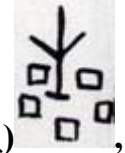
4. The principles of Oracle Bone selection

4.1. Distinction principles

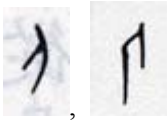
4.1.1. One or more types of components are different.

(莫)  ,  ，**(示)**  ,  ,  , 

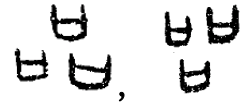
4.1.2. The number of components or lines is different.

(師)  ,  **, (星)**  , 

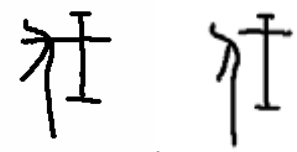
4.1.3. The direction (e.g. mirror image) of a component is different.

(人) 

4.1.4. The position of one or more components is different.

(品) 

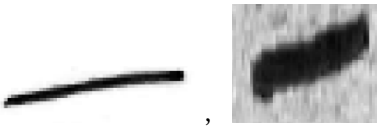
4.1.5. Whether the same set of components are connected each other or not.

(伐) 

4.2. Unification principles

Two or more instances of Oracle Bone with the following differences will be unified unless there's any difference in the meaning:

4.2.1. The length of corresponding line is different.

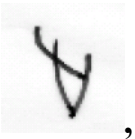
(一) 

4.2.2. The thickness of corresponding line is different.

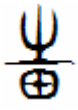
(王) 

4.2.3. The size of each component of the same set is different.

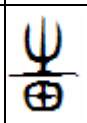
4.2.4. The enclosed part is filled or not filled.

(止)  , 

5. The principles of radical classification

5.1. If an Oracle Bone glyph is corresponding to Shuowen glyph, it should be classified into Shuowen radical of corresponded Shuowen glyph. For example,  (corresponded to ) should be classified to '明', not to '月' or '日'. In addition, Shuowen classifies some Guwen (古文) or Zhouwen (籀文) glyphs to a radical class even when these glyphs do not include Shuowen radical as their glyphic components. If the corresponded Shuowen glyph is such, the Oracle Bone glyph should be classified to the radical that the corresponded Shuowen glyph is included. For example,  (corresponded to ) is classified to '邑', not to '土' or '田'.

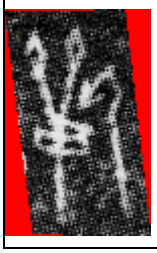
Original Script/Glyph	Corresp. S.W. Glyph	S.W. Radical

		明
		邑

5.2. The glyph should be classified into Shuowen radicals according to the Original Oracle Bone Inscriptions, e.g. "  " should be classified under radical 斤, because Shuowen radicals does not include the most significant glyphic component '單'. The next significant glyph component '斤' is used (KangxiZidian classifies '單' to the radical '口'. The classification of Oracle Bone shape "  " to '口' is more difficult to use than that to '斤').

歸部問題：以甲骨文字形爲主，如"  "入斤部。

Example 2

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

5.3. The shapes of the Original Oracle Bone inscriptions are different, but they share the same meaning and usage. Eventually, they have evolved into a pair of variants, with two different radicals in Shuowen. According to the radical classification in Shuowen, the

Oracle Bone glyphs are put under different radicals. e.g. 兀 and 元.
甲骨文異形同用，後世分爲兩字，說文分見兩部，則依《說文解字》收入不同部首。如“元”、“兀”。

Example 3

Imitation Script/Glyph	Original Script/Glyph	S.W. Radical
		一
		儿

5.4. The shape, meaning and usage of the Original Oracle Bone inscriptions are identical but new components have been added over time. If a radical can be found in Shuowen, the glyph will be put under the corresponding radical according to the Original Oracle Bone inscriptions. e.g. 彖 put under the 彖 radical and similarly 畐 in 畐 radical.
甲骨文同形同用，後世增添偏旁，《說文解字》另有部首者，則依甲骨文原形歸入相應部首。如“畐”入《說文解字》畐部，“彖”入《說文解字》彖部。

Example 4

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

		畐
		枭

5.5. The shapes of the Original Oracle Bone inscriptions are the same but they have many meanings and usages. Eventually, they have evolved into different characters. The glyph of these characters will be determined according to the shapes of the Original Oracle Bone inscriptions, and put under the corresponding radical in Shuowen. e.g. 史吏事.

甲骨文同形多用，後世分爲多字，依甲骨文字形分別隸定，歸入《說文解字》相應同一個部首。如：“史”、“吏”、“事”。

Example 5

Imitation Script/Glyph	Original Script/Glyph	S.W. Radical
		一
		一

6. **The principles of sorting the order of the glyphs of the same Oracle Bone Inscription**

同字之字形排序原則

6.1. Ordering of Glyph Categories

The Oracle Bone glyphs are classified into 3 groups; SW-mappable glyphs, SW-unmappable but with corresponded UCS character, SW-unmappable and without corresponded UCS character. They are ordered as following:

6.1.1. SW-mappable glyphs: The Oracle Bone glyphs that corresponded ShuoWen glyphs are identified are ordered by the order of ShuoWen Jiezi Daxu version (大徐本).

6.1.2. SW-unmappable glyphs with corresponded UCS character: The Oracle Bone glyphs that corresponded SW glyph is unavailable but corresponded UCS character is available should be placed after SW-mappable glyphs. To gather similar glyphs, the glyphs sharing same corresponded UCS characters should be collected to one group.

6.1.3. SW-unmappable glyphs without corresponded UCS character: The Oracle Bone glyphs that no corresponded SW glyphs and no UCS character should be placed after SW-unmappable glyphs with corresponded UCS character.

6.2. Ordering in Glyph Category

6.2.1. If one or more types of components or radical are different, then those with smaller difference will be placed first, and those with

greater difference behind.
異構字依字形差異大小排序，差異較小者置於前，差異較大者置於後。

6.2.2. Glyph variants will be placed after the typical glyph.
異寫字置於主形之後。

7. Release Process of the Database

For the record of the discussion of inclusion, deletion (because of unclear, cropped or exactly duplicated data), unification (submitted from different source but cannot be distinguished by the principles), or pending should be recorded in 'Status' column of the database.

8. Data Format For Oracle Bone Data Exchange

For the data exchange and review work, members are going to use the data format specified as follows:

- **Images format:**

8.1. use PNG storage format.

8.2. The original glyph should be scanned at 300 dpi (dots per inch).

8.3. The transcribed glyph images are named [ID]+[_R] (for example, if the ID is T00001, the transcribed glyph images should be named T00001_R).

8.4. The original glyphs are named [ID]+[_O] (for example, if the ID is T00001, the original glyph should be named T00001_O).

8.5. Glyph determination images are named [ID]+[_D] (for example, if the ID is T00001, the glyph determination image should be named T00001_D).

8.6. The images of unifiable shapes are named [ID]+[Unifiable shapes ID] (for example, if the Oracle Bone ID is T00001 and the unifiable shape ID is 000, the image of unifiable shape should be named T00001_000).

■ XML Schema:

```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema" elementFormDefault="qualified">
  <xs:element name="OldHanZi">
    <xs:complexType>
      <xs:sequence>
        <xs:element ref="Character" minOccurs="0" maxOccurs="unbounded"/>
      </xs:sequence>
      <xs:attribute name="version" type="xs:string" use="required" fixed="1.0"/>
    </xs:complexType>
  </xs:element>
  <xs:element name="Character">
    <xs:complexType>
      <xs:sequence>
        <xs:element ref="Source"/>
        <xs:element ref="Period"/>
        <xs:element ref="Area"/>
        <xs:element ref="Material"/>
        <xs:element ref="Radical"/>
        <xs:element ref="ModernChar" minOccurs="0"/>
        <xs:element ref="Unified" minOccurs="0" maxOccurs="unbounded"/>
        <xs:element ref="Note" minOccurs="0"/>
      </xs:sequence>
      <xs:attribute name="id" use="required">
        <xs:simpleType>
          <xs:restriction base="xs:string">
            <xs:pattern value="(G|T|K|KP|J|V|S|H|M)[0-9]+"/>
          </xs:restriction>
        </xs:simpleType>
      </xs:attribute>
    </xs:complexType>
  </xs:element>
  <xs:element name="Source" type="xs:string"/>
  <xs:element name="Period" type="xs:string"/>
  <xs:element name="Area" type="xs:string"/>
  <xs:element name="Material" type="xs:string"/>
  <xs:element name="Radical">
    <xs:simpleType>
      <xs:restriction base="xs:unsignedShort">
        <xs:minInclusive value="1"/>
        <xs:maxInclusive value="540"/>
      </xs:restriction>
    </xs:simpleType>
```

```
    </xs:element>
    <xs:element name="ModernChar" type="xs:string"/>
    <xs:element name="Unified">
      <xs:complexType>
        <xs:attribute name="id" type="xs:string" use="required"/>
      </xs:complexType>
    </xs:element>
    <xs:element name="Note" type="xs:string"/>
</xs:schema>
```

■ XML example:

```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<OldHanZi version="1.0">
    <Character id="T00001">
      <Source>甲骨文合集</Source>
      <Period>商</Period>
      <Area>河南安陽</Area>
      <Material>甲骨</Material>
      <Radical>001</Radical>
      <ModernChar>一</ModernChar>
      <Unified id="0000"/>
      <Unified id="0001"/>
      <Note/>
    </Character>
    <Character id="T00002">
      <Source>甲骨文合集</Source>
      <Period>商</Period>
      <Area>河南安陽</Area>
      <Material>甲骨</Material>
      <Radical> 005</Radical>
      <ModernChar>王</ModernChar>
      <Unified id="0000"/>
      <Note/>
    </Character>
</OldHanZi>
```

Universal Multiple-Octet Coded Character Set UCS

ISO/IEC JTC 1/SC 2/WG 2 IRG Nxxxx

Date: 2011-03-20

Source:	SUZUKI Toshiya, Faculty of Integrated Arts and Science, Hiroshima University
Title:	Japanese request of the update of Old Hanzi P&P.
Actions required:	For consideration at IRG #36
Distribution:	IRG Members and Old Hanzi Experts
Medium :	Electronic

By the minutes M35.8 agreed in IRG#35, Japan proposes to update Old Hanzi P&P described in IRG N1271 for further clarifications. There are 4 categories of the requests and proposals as follows:

1. Requests and proposals to clarify the input and the output of Old Hanzi ad-hoc group
2. Requests and proposals to clarify the glyph determination process.
3. Requests and proposals to clarify the glyph distinction rules.
4. Requests to reflect the revising that had ever agreed in previous meetings.

1. Requests and proposals to clarify the input and the output.

1.1 The reference books should be specified more precisely.

Some references are specified with the precise titles of the books, but some books are referred with abridged names, or with the names of the collections, of which the books are handling. For the convenience of the reviewers of the output of the ad-hoc group, the trackable reference information is essential. Following is the preliminary check result.

a) 甲骨文合集: “甲骨文合集” (13 volumes), 郭沫若主編, 中華書局, 1978-1982, ISBN 9787101016536. The earliest printings might not have ISBN numbers, but recent printing with exactly same content may have this number.

b) 甲骨文合集補編: “甲骨文合集補編” (7 volumes), 彭邦炯主編, 語文出版社, 1999, ISBN 7801264967?

c) 花園庄東地甲骨: book with exactly matching title cannot be found. It means “殷墟花園庄東地甲骨”, 中国社会科学院考古研究所, 雲南人民出版社, 2003, ISBN 9787222038776?

d) 山東済南大辛庄甲骨: book with exactly matching title cannot be found.

e) 周原甲骨: book with exactly matching title cannot be found. There is a book collecting the photographs of Oracle Bone in Zhou dynasty, titled "周原甲骨文", 曹瑋, 世界図書出版公司, 2002, ISBN 9787506256650. But its picture is insufficient to identify the glyphs.

f) 小屯南地甲骨: "小屯南地甲骨" (考古学専刊乙種 18 号), 中国社会科学研究院考古研究所, 中華書局, 1980(上冊) and 1983(下冊). ISBN is unidentified.

g) 英国所蔵甲骨集: "英国所蔵甲骨集" (上下), 李学勤, 中華書局, 1992, ISBN 9787101009569

h) 懷特氏等所藏甲骨文集: book with exactly same title could not be found.

i) 天理大学附属天理参考館蔵甲骨文: It means "ひとものこころ 天理大学附属天理参考館蔵品" Vol. 1, Num. 5, 道友社, 1986, ISBN 4-8073-0254-X?

j) 德瑞荷比所藏一些甲骨録: "Several Collections of Oracular Inscriptions in Germany, Switzerland, The Netherland, Belgium" (Fr. Jean Lefeuvre), Ricci, 1997, ISBN 9782950560247?

k) 瑞典斯德哥爾摩遠東古物博物館藏甲骨文字: "瑞典斯德哥爾摩遠東古物博物館藏甲骨文字", 李学勤, 中華書局, 1999, ISBN 7101022561?

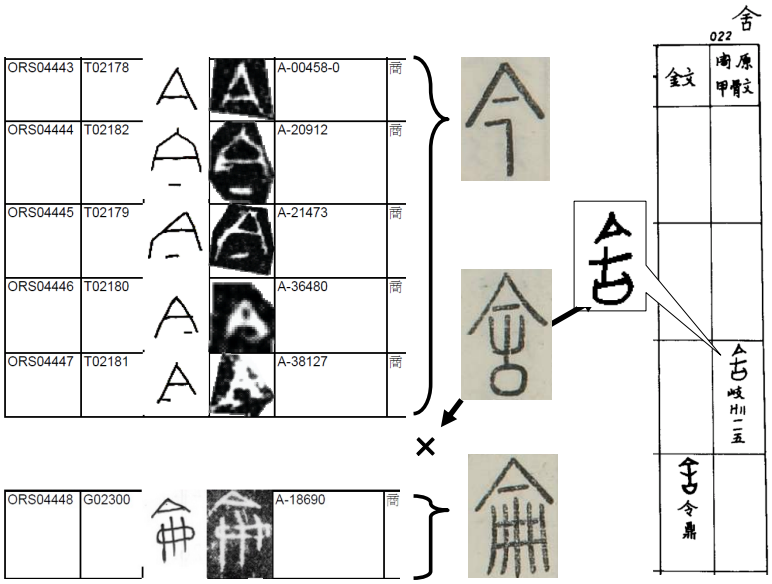
1.2 The coverage of the output glyph collection should be clarified.

In IRG N1168 (2005-11-29, IRG#25), China delegation had emphasized that the complete coverage is required to guarantee the multi use of encoded Old Hanzi, but it is unclear if it was agreed or not.

> 1. The principle of Old Hanzi Selection:
>
> The Old Hanzi should have a complete coverage and an exact map of the attributes of sources. Complete coverage is to guarantee the multiple use of Old Hanzi (research, publishing, virtual library).

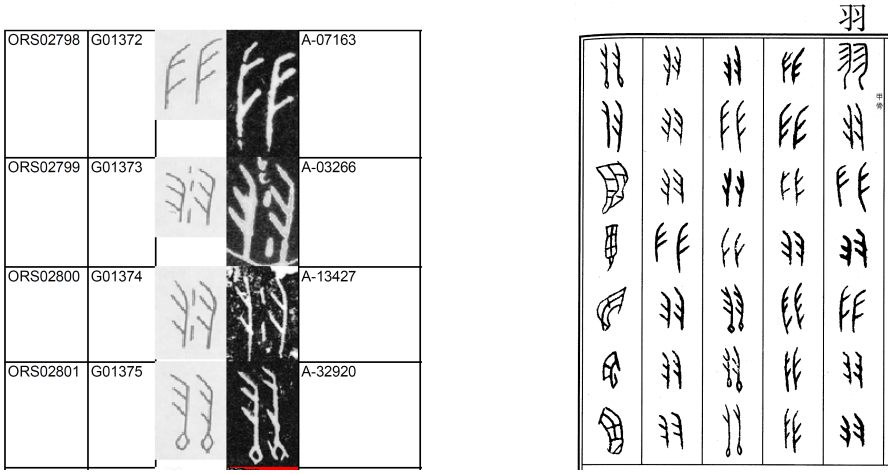
Chinese comment about the coverage, IRG N1168

The clarification of the coverage is important for the users of the coded character set to transliterate some Old Hanzi materials. If a user cannot find an Old Hanzi glyph in the coded character set, the user may request to add a *new* character. It is reasonable to add *new* character taken from newly-discovered materials (e.g. the excavation of the Oracle Bones in Zhou dynasty is still in progress). For example, "周原甲骨綜述" (徐錫臺, 三秦出版社, 1987?, ISBN 7-80546-062-0/K・26) shows an Oracle Bone glyph corresponding to "舍" which was not found in Shan-dynasty.



Missing Oracle Bone glyph “舍” in current database

But, Japan has a concern of the possibility that a user may find a new glyph from the references listed by P&P. For example, “古文字譜” (沈康年, 雲南人民出版社, 2006, ISBN 7-222-04695-X) shows several glyphs that should be distinguished by P&P but not included in current database. “古文字譜” does not provide the source information for each glyphs, so it is inappropriate to use it as the primal reference, but it is reasonable to use such dictionaries check the completeness of the current database.



The variations of Oracle Bone glyphs for “羽” in current database and in “古文字譜”.

So it is essential to have a reviewing period for the external experts who are interested in Old Hanzi to gather wider comments in accordance with more broad references. To minimize the review period, the list of the source glyphs that had ever been discussed in each contributing members or in the ad-hoc group would be required.

2. Requests and proposals to clarify the Glyph Determination process.

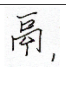
2.1 The coverage of the Kai style glyph for the replacement by meaning should be clarified.

As found in IRG#35, the determinated glyph in the current database may be resulted by the "replacement by meaning" rule proposed by TCA (see section 2 of IRG N1325, 2007-06-07).

To consider adding one principle regrding glyph determination (Liding) for Oracle Bone Inscriptions to clarify how the glyph determination (Liding) should be made when the shape of Liding cannot be decided based on the Original Script/Glyph of the Oracle Bone Inscriptions, but rather based on the context**"**. Details of this principle will be finished at the next Old Hanzi Group meeting.

甲骨文原形無法據其形隸定者，其隸定原則於下次會議討論。

Example 1: Glyph determination according to the original script is not possible 無法據其形隸定

Original Shape/Glyph	Rep. Script/Glyph	Glyph Determ. (Liding)
		
		

Example 2: Glyph determination according to the original script is possible 可據其形隸定

Original Shape/Glyph	Rep. Script/Glyph	Glyph Determ. (Liding)
		

Rule to replace by meaning rule proposed by TCA (IRG N1325)

However, many dictionaries had ever invented to simulate the structures of Old Hanzi, it is difficult to decide if the appropriate Kai style is really unavailable or not. For example, CJK Unified Ideographs Extension B includes many forcibly transformed Kaishu of Small Seal characters via KangxiZidian. Also there are several publishing that includes the forcibly transformed Kaishu of Old Hanzi (e.g. "古文字詁林" have developed yet-another transformations of Shuowen glyphs).

 →  (coded at U+2666A)

Example of Kaishu that is forcibly transformed from ZhuanWen

It is impossible to collect all possibilities, so the artificial restriction will be useful to

restrict the coverage of possible Kai style glyph components by the collection of the authorized dictionaries, or the current coded character set.

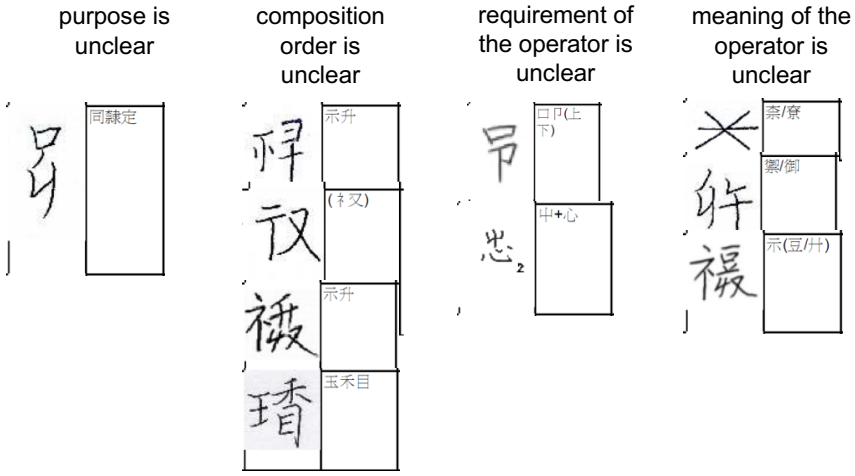
2.2 The syntax of “Corresponding Modern Character” column should be clarified.

In current P&P, the corresponding modern character is defined as follows.

> 3.5 If the field “Corresp. Modern Char.(UCS Code)” is blank, then the “Notes field” must be filled with justifications to indicate the glyph is well-understood, for example, the meaning of the ‘Rep. Script / Glyph’.
>
> 3.6 For the field ‘Corresp. Modern Char.(UCS Code)’ there may be no corresponding Simplified character only corresponding Traditional characters.

The definition of “Corresp. Modern Char.” column in current P&P

The rule 3.5 can be understood as the rule to fallback replacement by the meaning, and the rule 3.6 can be understood as the rule to harmonize the efforts by China and TCA. Anyway, as the title of the column tells, the column should include 1 UCS character or nothing. But the current database uses the column to note a memo about the structure by unclear syntax.



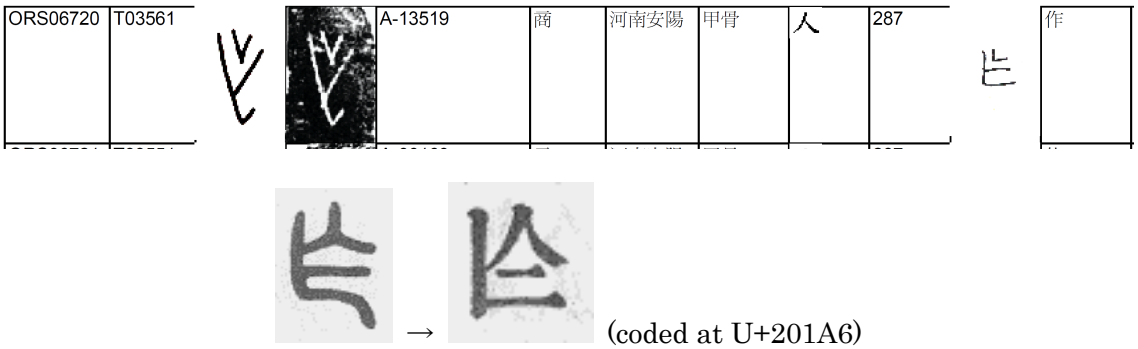
Examples of inconsistent usage of “Corresp. Modern Char.” in current database.

Japan suggests updating the entries to fit the original definition, 1 UCS character or leaving blank (if no appropriate UCS is available). However, considering that TCA database might include the glyph described by the string (see IRG N1165 Fig. 2), there is a room to discuss to add a new column to store such information.

If the description by the strings of UCS is required for any purpose (for example, the algorithmic sorting of the data by using the data in this column), the syntax and the possible components should be clarified. As found in the example, some subcontractors used CJK Radical “礻” (mainly G data) and others used CJK Unified Ideograph “示”

(mainly T data), they are difficult to support the machinery processing of the database.

In addition, the coverage of the UCS characters for this column is unclear. For example, the Kaishu-fied Old Hanzi for “乍” is already coded via Kangxi Zidian source (the character is included in both of GB 18030 and CNS 11643), but the current database does not use it and “作” is assigned.



(coded at U+201A6)

Example of overlooked UCS character fitting to the determinated glyph

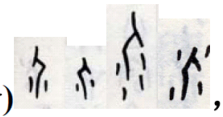
From the observation of the working process, Japan has a concern that UCS is too large collection for Old Hanzi experts, and proposes to investigate the existing typesetted dictionaries (e.g. 新甲骨文編) published in each national body and define a compact subset of UCS for the database working in progress.

3. Requests and proposals to clarify the glyph distinction rule.

3.1 The method to count the line number should be clarified, with better examples.

In the rule of 4.1.2, the number of components and lines are focused. But the examples for “介” category includes the glyphs with same number of lines but the direction is different, so this example does not match with the title.

4.1.2 The number of components or lines is different.

(介)  , (星) 

Examples to show the different numbers of component and stroke

The examples for “星”, the 1st and 3rd glyphs are difficult to determine if the number of stroke is same or not. “V” is counted as 2 lines and “U” is counted as 1 line?

Most modern Hanzis are drawn by standardized strokes, so the most dictionaries omit the definition of the stroke counting, but the structure of Oracle Bone character is different. Japan propose to remove the 2nd and 4th glyphs for “介”, and the 3rd glyph for “星”. The 2nd glyph of “星” is questionable too. It is unclear if the lowest horizontal line of the 1st glyph of “星” is recognized as an independent component, or recognized as a serif-like part of the vertical line. Japan requests to add some document to clarify the method of counting number of components or lines.

3.2 The classification of the component positions should be clarified.

The relative positioning of the components in modern Hanzis are well patternized as IDC was defined, but it is not applicable to Old Hanzi.

4.1.4 The position of one or more components is different.

(好) 

Examples to show the different position relationship

For example, the 1st, 4th and 5th glyph for “好” category share similar position relationships. If 1st and 4th are distinguished as “1st is left-and-right, but 4th is left-upper-and-right-lower”, how the 5th glyph should be distinguished? The granularity of the component position classification should be clarified, as IDC specifies it for modern Hanzi. Japan proposes to remove the 3rd, 4th and 5th glyphs for “好” and giving the list of the classification of the relative positions.

3.3 The rule to distinguish the stroke is followed in current works?

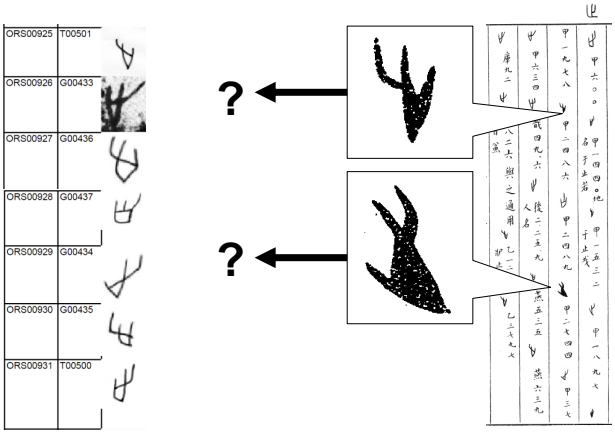
The rule 4.1.6 requests very fine granularity to distinguish the glyphs.

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

(王) 

Examples to show the different stroke types

Especially, the distinction between 5th and 6th glyphs is too sensible. Following is the comparison of "止" glyphs in the current database and those in "甲骨文編".



Examples of filled variants for "止" in 甲骨文編

There are 2 "filled" variants of "止" that are not included in the current database. Japan has a concern that the subcontractors were not aware of the rule. To avoid the long revisiting of the glyph database, Japan suggests removing this rule by following reasons:

- In the case of Oracle Bone script, the variant glyphs with filled area are popular for a small group of the glyphs, like, 王, 月, 止, etc. However, the filling in Bronze inscriptions is more popular, it should be discussed in the working for Bronze inscriptions.
- No examples are known that shows semantic or phonetic differences by the filling or unfilling.
- It is difficult to reflect the filling or unfilling in determinated glyph.

4. Requests to include the previously discussed and agreed updates.

4.1 Definition of Shuowen order (raised by TCA in IRG#24)

In IRG N1123 (2005-05-17), TCA had proposed to use Shuowen Jiezi Zhu by Duan-Yu-Cai-Zhu.

> 5. What radical should be taken?
>
> TCA propose:
>
> Take the radical order based on Duan-Yu-Cai-Zhu of Shuowen Jiezi Zhu(footnotes) by Xue-Shen as a rule.

TCA proposal to clarify the Shuowen

However, current P&P is still unclear which SW is used (see 2.4 “SW Radical” and 2.5 “SW Radical number”). If there is an Oracle Bone glyph corresponding to the additional glyph (新附字) in 大徐本, fixing a version of Shuowen is essential to determine the position to place the character. If 大徐本 is used, these glyphs should be placed before the glyphs that Shuowen has no corresponding glyphs.

In addition, the sorting order of the glyphs under the same radical is unclear in current P&P. In the section 6 “The principles of sorting the order of the glyphs of the same Oracle Bone Inscriptions”, there is no rule to determine the order of the glyphs that Shuowen includes exactly corresponding character. So it is difficult to understand why “吏” should be placed before “天”, why “帝” and “旁” should be placed between “上” and “下”. Fixing a concrete edition of Shuowen Jiezi and the clarification of the order of each glyph is expected.

Also Japan proposes to add new column to note the temporal order of the glyph in the database. At present, the order of the glyphs is manually corrected in each face-to-face meeting, but it is difficult to summarize the correction of the order. It is because current database has no persistent entry to store the order of the glyphs, thus we cannot compare the order before the meeting and after the meeting. The column for the temporal order is useful not only for the tracking the change but also for the submission of the ordering update proposal by the electronic document. It will accelerate the speed of the database correction, and the experts can use the time of face-to-face meeting for more detailed discussion that requires their expertise.

4.2 Yin-Gou and Yi-Xie (raised by TCA in IRG#24)

In IRG N1123 (2005-05-17), TCA had proposed to add a rule to distinguish Yi-Gou and Yi-Xie. TCA commented that Yi-Gou should be coded but Yi-Xie should not be coded. In the meeting report IRG N1135, it was concluded that Yi-Gou feature should be noted. But current P&P does not mention about Yi-Gou and Yi-Xie issues, it should be clarified.

Also Japan proposes to use appropriate English words to describe Yi-Gou and Yi-Xie issues.

4.3 Contextual Information (raised by China in IRG#25)

In IRG N1168 (2005-11-29), China had listed 2 information to be collected in the Old Hanzi database; contextual information and glyph information.

2. Procedures:

To facilitate the processing of Old Hanzi, a database should be built. It should have the following platforms:

A. Contextual information

The text in which the Old Hanzi appears, the attribute of region, time, scholarly research work, and other related information. (see fig. 1)

B. Glyph information

The original image, decomposing component, representative glyph, corresponding modern character (if any), coverage index of the selected glyph. (see fig. 2)

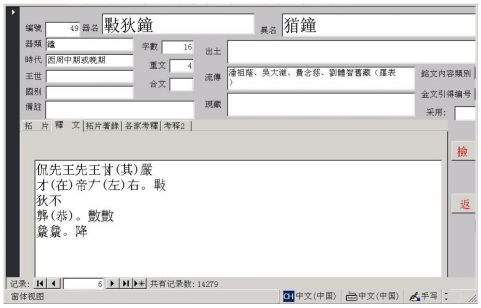
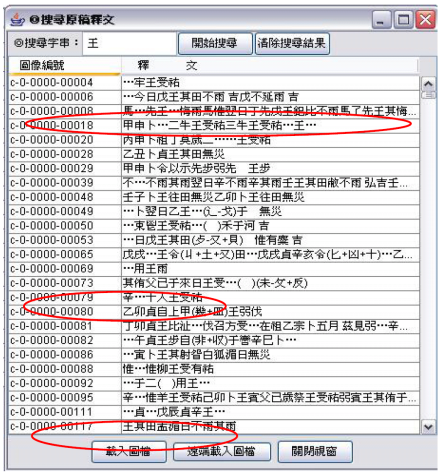


Figure 1

Chinese suggestion about the contextual information (IRG N1168)

TCA didn't mention about the requirement of the contextual information, but the existing database reported in IRG N1165 (2005-11-29) included such.

Fig. 2 the frame of using explanation archives search



TCA database for contextual information (IRG N1165)

According to the meeting report of IRG#35, IRG N1182, it was agreed to collect both information (see meeting report IRG N1182, 2-(2)).

2. The consolidated Old Hanzi database will include:

(1).The original images of the Old Hanzi and the explanation archives.

(2) The glyph determination and corresponding modern character.

(3) The attribute data of collected Old Hanzi inscriptions.

(Please refer to database illustrations in IRG N 1165 and N 1168.)

Agreement about the database in IRG#25

But current Old Hanzi database seems to hold only glyph information, no contextual information. It is expected that the contextual information are not created for this project from scratch, but based on existing studies like “甲骨文合集釈文”. There are some copyright issues to redistribute them their content via the database, so it is expected what source was used in the database.

4.4 Revise of the principles of Old Hanzi selection (agreed in IRG#30, 2007)

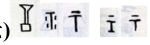
According to IRG N1460, it was agreed to revise the principles of Old Hanzi selection, to add new data to the database (see 3.1 and 3.2 of IRG N1460); a representative glyph should be chosen for a category of the glyphs sharing same meaning and pronunciations (so, this “representative glyph” is per-category glyph, not per-entry data). The agreed revise should be executed.

3.1 Revise the principles of Old Hanzi selection 修訂選字原則

Characters with the same meaning and pronunciation but with different structure or component will be put into the same category. A representative glyph will then be chosen.

字形之分類概念，將音義相同，字形構字或部件不同的字歸為一類，並選取一個字作為代表字。

Example:

(莫)  ，(示) 

3.2 The principles of selecting a representative glyph:

代表字的選取原則

Based on the origin shape of the Oracle Bone Inscriptions, the most representative glyph will be selected.

以甲骨文原形來做判斷，選取最有代表性之字形。

Example: use the first glyph as the representative (such as )

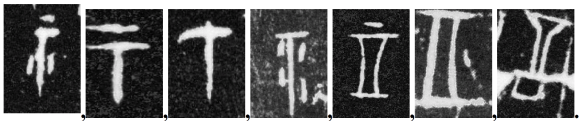
範例：以第一個字為代表字(如 )

(莫)  ，(示) 

The agreed update of P&P in IRG#31 (IRG N1460)

China has ever raised similar proposal for the case that one character have multiple glyphs. China proposed to select a glyph that is most similar to Shuowen glyph, but Japan considers that it is arguable for the cases that the glyph different from Shuowen is most popular in Oracle Bone objects.

(2) If one character has more than one glyph , the sequence of these glyph in the list is that the glyph of same with or similar to the small seal of Shuowen Jiezi is arranged in the first, others are arranged according to the principle of similarity. For example, the glyph of the character "示" are arranged as follow:



IRG N1261 (2006-11-26, raised by China)

Current database uses "representative glyph" in different meaning. In the original discussion about the "representative", IRG N1087 (TCA, 2004-11-07), "representative" is used to keep the basic structure and exclude minor writing variants of the strokes.

2.1.4 Writing-variances: Variant constructs for a character need to be encoded separately. In the case of variant writing, the representative writing should be chosen. Other writing-variances should be listed aside with the representative. Characters with variant constructs may be listed as notes. A construct-variance replaces a component in a character with a different component or a component with different writing. The writing-variance involves some stroke(s) in a character having minor different drawing(s), however the basic structure of the character is retained. The evolution of old Hanzi took a long time. The decision involves the structure differences of basic strokes. For example, if a dot in a character became a horizontal bar, the character is deemed as a construct- variance. However, characters with different sizes of dot of different lengths of horizontal bar will make them writing-variances.

3.1.1 Representative characters: Representative characters were chosen by experts. Rules for collecting representative characters are as follow: well-recognized characters are collected; uncertain or arguable characters are left out. Representative characters are manually drawn careful with aesthetic and uniform styles. Original scripts are scanned into graphic files with clear resolutions. Please see the appendix: Ancient Jinwen script encodings table.

The earliest TCA definition of the "representative" (IRG N1087)

But the definition of "Rep. Script/Glyph" in current P&P is quite different. It is defined as "The truthful trace from ‘`Original Shape/Glyph"'. And, the usage of "Rep.

Script/Glyph" column is not consistent; some subcontractors (mainly for G data) make a digitally negated picture from original gray-scale picture, other subcontractors (mainly for T data) draw the manual copies of the glyphs. It does not *represent* the glyph but show a glyph instance, because the picture at the column does not exclude the writing variance.

Respecting the long work of current database, Japan proposes to add "referential glyph" for the revise agreed in IRG N 1460.

4.5 Revise of the radical classification (agreed in IRG#30, 2007)

According to IRG N1460, it was agreed to revise the principles of radical classification to improve the similarity with Shuowen (see 3.3 of IRG N1460); an Oracle Bone should be classified to a Shuowen radical that includes a corresponding glyph, even if the Oracle Bone glyph does not include the radical as its glyphic component. Some 籀文 or 古文 glyphs don't share the radical with the related heading character, but they are classified in the same radicals. To reduce the ambiguity, Japan requests to update P&P to include this rule agreed in IRG N1460.

3.3 Principle of radical classification

歸部原則

For a character that is in Shuowen, the same radical as in Shuowen will be used. For a character that is not in Shuowen, the most suitable component in it will be chosen as the radical, on the condition that the radical chosen is also in Shuowen.

若有對應說文的字，則依照說文歸部;如不包含說文部首之形，則依照該字形適當之偏旁附註部 2。

For example 1:

 說文《  部》有對應字的重文『明』，則  歸部。

For example 2:

 說文《邑部》有對應的『邦』字的重文  ，則 

歸邑部，並在附註中註明部首 2 田部。

4.6 New column to indicate the status of unified or deleted (agreed in IRG#35, 2011)

As the meeting report IRG N1746 notes (2.2 and 2.3), Old Hanzi ad-hoc group agreed to add new column “status” to indicate the deleted, kept or unified glyphs.

(end of document)

ISO/IEC JTC1/SC2/WG2/IRG N1747_KimKFeedback
Doc. #: Korea JTC1/SC2 K2060

Korea JTC1/SC2, Committee on Character Codes

KIM, Kyongsok, Chairperson of Korea JTC1/SC2
email: gimgs0 AT gmail DOT com, phone: +82-51-510-2292
address: Division of Computer Science and Engineering,
Pusan National University
2 Busandaehagro 63 Beon Gil, Geumjeonggu, BUSAN 609-735, Rep. of KOREA

Author: KIM, Kyongsok
Date: 2011.06.16.
Status: Individual Expert's contribution
Subject: comments RE: IRG N1747 Old Hanzi P&R v3
Relevant documents: IRG N1747

The author would like to make some comments RE: IRG N1747 Old Hanzi P&R (v3). The author hopes that appropriate comments be adopted and reflected in the next version of Old Hanzi P&R.

When a rationale seems obvious, it is not shown.

Since the author is not an expert of Old Hanzi, the author did not try to change the important concepts. Instead, the author tried 1) to make P&R more readable and understandable; and 2) to use terms consistently.

1. p. 1, 1. Format of submission

1. Format of submission

ID	Imitated Script/Glyph	Original Shape/ Glyph	Source	Period/ Epoch	Area /Site	Material	SW Radical	SW Radical Number	* Corresp. Modern Char (UCS Code)	* Unifiable Shapes	* Notes

1.1 Imitated Script/Glyph
-->
Imitated Glyph

= Rationale: Glyph and Script do not seem to be interchangeable.
  Glyph seems appropriate.

= The same comment applies to: p. 2, 2.2, 3.1 & 3.2;
   p. 8, 5.2; p. 9, 5.3 & 5.4; p. 10, 5.5

1.2 Original Shape/Glyph (sometimes Original script/glyph)
  -->
    Original Shape/Glyph (sometimes Original glyph)

= Rationale: .
  - To be consistent with the previous column "Imitated Script/Glyph", using both "Imitated Glyph" and "Original Glyph" seems appropriate.
  - On p. 2, in 2.3, "Original glyph", NOT "Original Script/glyph", is used.

= The same comment applies to p. 7, 5.1; p. 8, 5.2; p. 9, 5.3 & 5.4;
    p. 10, 5.5

1.3 (UCS Code)
  -->
    (UCS CodePos)

= Rationale: The term "UCS CodePos" clearly identifies what to put in this column. "UCS Code" may be somewhat confusing.

= The same comment applies to: p. 3, 3.5 & 3.6;

1.4 "SW" vs. "S.W."

  - Here "SW" is used.
  - For example, at the bottom of p. 7, "S.W." is used.

  - The author suggests to use "SW" or "S.W." consistently.

2. p. 1, 2. Definitions

2.1. ID: It is the unique id that consists of one or two letters member id (G, T, K, KP, J, V, S, H, M) followed by four digit sequential numbers assigned by submitters.

2.1 one or two letters member id

-->
one or two letter member id

2.2 four digit sequential numbers

-->
a four digit sequential number

3. p. 2, 2.1

Example: T0001 is one IRG global unique ID assigned to an Oracle Bone submitted by TCA.

3.1 to an Oracle Bone submitted
-->
to an Oracle Bone character submitted

= The same comment applies to: p. 2, 2.3

4. p. 2, 2.4

provide the picture, it is assigned ...
-->
provide the picture; it is assigned ...

5. p. 2, 3.1

= What is meant by "standardized size and dimension"?
What is the difference between size and dimension?

= The author suggests to add some explanation to this phrase.

6. p. 2, 3.4

The last three columns are ‘optional’ field and they are indicated with an asterisk “*”, the other fields are mandatory.

-->

The last three columns are ‘optional’ ~~field~~ and they are indicated with an asterisk “*”; the other ~~fields~~ cloumns are mandatory.

7. p. 3, 3.5.2

Only URO (CJK Unified Ideographs in BMP without Extension A and CJK Compatibility Ideograph)

-->

Only CJK Main (CJK Unified Ideographs in BMP excluding CJK Extension A and CJK Compatibility Ideographs)

= Rationale: URO is not defined in ISO/IEC 10646.
  - Furthermore, URO is an acronym of Unified Repertoire and Ordering, which means that URO refers to not only repertoire but also ordering (Plz correct him if the author is wrong).
    Therefore, URO seems somewhat inappropriate to refer to only a repertoire (CJK Main).
  - CJK Main is not defined in ISO/IEC 10646 either.
  - However, CJK Main is sometimes used informally.

  - Probably we need to add a term to ISO/IEC 10646 which refers to only CJK Main.

8. p. 3, 3.6

to indicate the glyph is well-understood, for example, the meaning of the ‘Imitated Script/Glyph’.

-->

to make the glyph well understood (for example, the meaning of the glyph).

9. p. 3, 3.7

9.1
The “Source” field consists of two data elements with one optional element.

-->

The “Source” field consists of two mandatory ~~data~~ elements and one optional element.

9.2
The 1st letter indicates the book reference number

-->

The 1st element is a book reference letter.

= Rationale:
To be consistent, explanations begin with "the n-th element is ...".

10. p. 5, the first two bullets are duplicated. The two lines in the second bullet need to be deleted.

- (mandatory) Oracle Bone number (甲骨拓片的編號) which consists of 5 digits assigned uniquely to each Oracle Bone.

- (mandatory) Oracle Bone number (甲骨拓片的編號) which consists of 5 digits assigned uniquely to each Oracle Bone.

11. p. 5, the first bullet

11.1

● (mandatory) Oracle Bone number (甲骨拓片的編號) which consists of 5 digits assigned uniquely to each Oracle Bone.

-->

● (mandatory) The 2nd element is a Oracle Bone character number (甲骨拓片的編號) which consists of 5 digits assigned uniquely to each Oracle Bone character.

= Rationale:
  1) See rationale for 9.2
  2) Oracle Bone character clearly indicates one character.

11.2 Question. Is this 5-digit number unique 'within each book' or unique 'among all books'?
  - For example, can we have both A-00001 and B-00001?
    Or if we have A-00001 then we cannot have B-00001 but can have B-00002 instead?
  - Depending on the meaning of uniqueness, the above statement in first bullet may have to be elaborated appropriately.

12. p. 5, second bullet ('second' after deleting one duplicated bullet)

● (optional) Identifier to determine the side of the Oracle Bone which consists of 1 digit. The possible values are '0' for front side, '1' for back side. If an inscription is carved only on one side, this element will be omitted.

-->

● (optional) The 3rd element indicates the side of the Oracle Bone character which consists of 1 digit. The possible values are '0' for front side, '1' for back side. If a character is carved only on one side, this element will be omitted.

= Rationale:
  1) See rationale for 9.2
  2) Oracle Bone character clearly indicates one character.

13. p. 5, 4.1

Distinction principles

-->

Separation principles

= Rationale:
  - Since we want to code two (or more) glyphs separately, "separation" seems more appropriate than "distinction".

14. p. 5, Add the following statement between 4.1 and 4.1.1

[none]

-->

Two or more instances of Oracle Bone characters with the following differences will be separated:

= Rationale: To be consistent with 4.2 below.

15. p. 5, 4.1.1

One or more types of components are different.

-->

One or more components of characters are different.

16. p. 6, 4.1.5

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

-->

Connectivity of components is different.

= Rationale: To be consistent with 4.1.1 ~ 4.1.4.

17. p. 6, between 4.2 and 4.2.1

Oracle Bone

-->

Oracle Bone character

= Rationale:
  - Oracle Bone character clearly indicates one character.

18. p. 6, 4.2.3

The size of each component of the same set is different.

-->

The size of corresponding components is different.

19. p. 7, 4.2.4

The enclosed part is filled or not filled.

-->

The enclosed part of one character is filled and the corresponding part of another character is not filled.

= Rationale: To be consistent with 4.1.1 ~ 4.1.4 and 4.2.1 ~ 4.2.3

20. p. 7, 5.1

If an Oracle Bone glyph is corresponding to Shuowen glyph

-->

If an Oracle Bone glyph corresponds to Shuowen glyph

21. p. 7, 5.1; Use "classified into" consistently.
  - This comment applies to several places on pp. 7 ~ 8.

should be classified into Shuowen radical ...
...
should be classified to ‘明’, not to ‘月’ or ‘日’.

-->

should be classified into Shuowen radical ...
...
should be classified into ‘明’, not into ‘月’ or ‘日’.

22. p. 7, 5.1, Use "corresponding" consistently.
  - This comment applies to several places on pp. 7 and 11.

corresponded Shuowen glyph

-->

corresponding Shuowen glyph

23. p. 8, 5.2

Original Oracle Bone Inscriptions

-->

Original Oracle Bone characters

= Do we need to use the term "inscription" here? If so, we need to define "inscription" and to explain cleary why inscription is better than character.

23. p. 8, 5.2

because Shuowen radicals does not include ...

-->

because Shuowen radicals do not include ...

24. p. 7, 5.1
  - There are Examples 2, 3, 4, and 5.
  - But there is no Example 1. ??
  - Probably we need to add "Example 1" for the table at the end of 5.1.

25. p. 8, 5.3

The shapes of the Original Oracle Bone inscriptions are different, but they share the same meaning and usage. Eventually, they have evolved into a pair of variants, with two different radicals in Shuowen. According to the radical classification in Shuowen, ...

-->

Suppose that the shapes of the Original Oracle Bone ~~inscriptions~~ characters are different, but that they share the same meaning and usage. Eventually, they have evolved into a pair of variants with two different radicals in Shuowen. In such a case, according to the radical classification in Shuowen, ...

26. p. 9, 5.4

The shape, meaning and usage of the Original Oracle Bone inscriptions are identical but new components have been added over time. If a radical can be found in Shuowen, the glyph will be put under the corresponding radical according to the Original Oracle Bone inscriptions. e.g. 彔 put under the 彔radical and similarly 畐 in 畐 radical.

-->

Suppose that the shape, meaning and usage of the Original Oracle Bone ~~inscriptions~~ characters are identical but that new components have been added over time. If a radical can be found in Shuowen, the glyph will be put under the corresponding radical according to the Original Oracle Bone ~~inscriptions~~ character. e.g. 彔 put under the 彔 radical and similarly 畐 ~~in~~ under 畐 radical.

27. p. 10, 5.5

The shapes of the Original Oracle Bone inscriptions are the same but they have many meanings and usages. Eventually, they have evolved into different characters. The glyph of these characters will be determined according to the shapes of the Original Oracle Bone inscriptions, and put under the corresponding radical in Shuowen.

-->

Suppose that the shapes of the Original Oracle Bone ~~inscriptions~~ characters are the same but that they have many meanings and usages. Eventually, they have evolved into different characters. The glyph of these characters will be determined according to the shapes of the Original Oracle Bone ~~inscriptions~~ characters and put under the corresponding radical in Shuowen.

28. p. 11, 6., The title of Section 6.

The principles of sorting the order of the glyphs of the same Oracle Bone Inscription

= Since the author is not an expert of Old Hanzi, the author cannot understand the meaning of title of Section 6.

The author suggests that
1) examples can be added to each of 6.1.1, 6.1.2, 6.1.3, 6.2,1 and 6.2.2;
2) clarify the difference between 6.1 and 6.2; and
3) if appropriate, add 6.3 which explains ordering using examples when both 6.1 and 6.2 are combined.

29. p. 11, 6.1

The Oracle Bone glyphs are classified into 3 groups; SW-mappable glyphs, SW-unmappable but with corresponded UCS character, SW-unmappable and without corresponded UCS character. They are ordered as following:

-->

The Oracle Bone glyphs are classified into 3 groups: 1) SW-mappable glyphs, 2) SW-unmappable glyphs ~~but~~ with corresponding UCS character, and 3) SW-unmappable glyphs ~~but~~ without corresponding UCS character. They are ordered as follows:

= A similar comment applies to 6.1.1, 6.1.2 and 6.1.3.

30. p. 11, 6.1.1

= The following statement is confusing.

The Oracle Bone glyphs that corresponded ShuoWen glyphs are identified are ordered by the order of ShuoWen Jiezi Daxu version

-->

The Oracle Bone glyphs that correspond to ShuoWen glyphs are identified ~~are~~ and ordered by the order of ShuoWen Jiezi Daxu version

= The author is not sure if his suggested statement correctly reflects the intent of the original statement.
= The author suggests to improve the statement.

31. p. 11, 6.1.2

the glyphs sharing same corresponded UCS characters should be collected to one group.

-->

the glyphs sharing the same corresponding UCS characters should be collected ~~to~~ as one group.

= The author is not sure if his suggested statement correctly reflects the intent of the original statement.

32. p. 11, 6.1.3

SW-unmappable glyphs without corresponded UCS character: The Oracle Bone glyphs that no corresponded SW glyphs and no UCS character should be placed after SW-unmappable glyphs with corresponded UCS character.

-->

SW-unmappable glyphs without corresponding UCS character: The Oracle Bone glyphs for which there is neither corresponding SW glyphs nor UCS character should be placed after SW-unmappable glyphs with corresponding UCS character.

33. p. 12, 7.

33.1

unification (submitted from different source but cannot be distinguished by the principles),

-->

unification (glyphs that came ~~submitted~~ from different sources (see 3.7) but that cannot be ~~distinguished~~ separated by the separation principles),

33.2
= Question: Does "source" refer to "member body" or "book reference"?
  - It seems to refer to book reference.

34. p. 12, 8.

34.1 8.
Data Format For Oracle Bone Data Exchange

-->

Data Format For Oracle Bone character Data Exchange

34.2 p. 12, 8.3

[ID]+[_R]

-->

[ID] + "_R"

34.3 p. 12, 8.3

  = The author suggests that definition of "transcribed glyph image" be added.

34.4 p. 12, 8.4

[ID]+[_O]

-->

[ID] + "_O"

34.5 p. 12, 8.5

[ID]+[_D]

-->

[ID] + "_D"

34.6 p. 12, 8.5

Glyph determination images are named ...

= "Glyph determination images" does not seem to have been defined.
= The author suggests that definition of "Glyph determination images" be added and, if appropriate, the above statement be modified accordingly.

34.7 p. 12, 8.6

[ID] + [Unifiable shapes ID]

-->

[ID] + "_" + [Unifiable shapes ID]

34.8 p. 12, 8.6

if the Oracle Bone ID is T00001 ...

-->

if the ~~Oracle Bone~~ ID is T00001

34.9 p. 12, 8.6

= Since there is already "ID", "Unifiable shapes ID" could be confusing.
= The author suggests that "Unifiable shapes ID" be changed to some other term (e.g., "Sequence number of a Unifiable shape")

= The author is not sure if his suggested statement correctly reflects the intent of the original statement.

41. There are no page numbers in P&R v3.
- The author suggests to add page numbers in the future versions of P&R.

42. The author wonders why Old Hanzi uses P&R, not P&P.
- Sometimes in the IRG doc. list, Old Hanzi P&P, NOT P&R, is used which causes confusion.
- Unless there is a definite reason to use P&R instead of P&P, The author suggests to change P&R to P&P in the future versions of P&R/P&P.

* * *