Technical Reports

L2/08-284

Draft Unicode Technical Report #45

U-SOURCE IDEOGRAPHS

Author	John Jenkins 井作恆 (jenkins@apple.com)
Date	2008-06-04
This Version	http://www.unicode.org/reports/tr45/tr45-1.html
Previous Version	n/a
Latest Version	http://www.unicode.org/reports/tr45/
Tracking Number	1

Summary

This document describes U-source ideographs as used by the Ideographic Rapporteur Group (IRG) in its CJK ideograph unification work.

Status

*This is a **draft** document which may be updated, replaced, or superseded by other documents at any time. Publication does not imply endorsement by the Unicode Consortium. This is not a stable document; it is inappropriate to cite this document as other than a work in progress.*

> ***A Unicode Technical Report (UTR)** contains informative material. Conformance to the Unicode Standard does not imply conformance to any UTR. Other specifications, however, are free to make normative references to a UTR.*

Please submit corrigenda and other comments with the online reporting form [Feedback]. Related information that is useful in understanding this document is found in References. For the latest version of the Unicode Standard see [Unicode]. For a list of current Unicode Technical Reports see [Reports]. For more information about versions of the Unicode Standard, see [Versions].

Contents

1 Introduction

This document describes U-source ideographs as used by the Ideographic Rapporteur Group (IRG) in its CJK ideograph unification work. The IRG is a subgroup of ISO/IEC JTC1/SC2/WG2 and has the formal responsibility of developing extensions to the encoded repertoires of unified CJK Ideographs. The IRG consists of members of ISO/IEC member bodies and liaison organizations, including many East Asian countries and the USA. The Unicode Consortium participates in this group as a liaison member of ISO.

The U-source consists of the CJK ideographs which have been submitted to the UTC as potential candidates for encoding. Not all of these are, in fact, suitable candidates for encoding, and their inclusion in this document should not be taken as approval for their encoding on the part of the

UTC.

The actual U-source data are found in two additional files:

- [Glyphs], a PDF showing the glyphs for the U-source ideographs.
- [Data], a text file containing information regarding the ideographs.

2 Text File Data

The text file consists of UTF-8 text. Each line consists of seven fields separated by semicolons.

1. The ideograph's U-source identifier. This consists of the letters "UTC" followed by five decimal digits, starting with 00001. Identifier numbers are not skipped, and are not reused. Identifier numbers are assigned sequentially.
2. A single character indicating the ideograph's current status. These are described below.
3. A Unicode code point. This field is empty if the status is not C, U, or V. The meaning of this field in these three cases is described below.
4. A radical-stroke index for the ideograph, as described in [UAX38].
5. A KangXi dictionary index for the ideograph, as described in [UAX38].
6. An ideographic description sequence (IDS) for the ideograph, if one can be generated.
7. A string indicating the ideograph's source and an optional index within the source.

2.1 The Status Field

The status field reflects the ideograph's current status. The value of this field can change over time. The possible values are C, D, E, U, V, W, and X; new values may be added in the future.

A status of C means that the ideograph is found in Extension C. This is currently under ballot in WG2. The Unicode field here indicates the proposed code point being ballotted.

A status of D means that the ideograph has been submitted to the IRG as part of the UTC's Extension D proposal.

A status of E means that the ideograph has been submitted to the IRG as part of the UTC's Extension E proposal.

A status of U means that the ideograph is already encoded in Unicode. Characters with a status of U were either added to the U-source database in error, or are characters encoded in Unicode before the IRG began its work. The Unicode field here is the code point for the encoded character.

A status of V means that the ideograph is a variant of a character encoded in Unicode. These variants are not limited to Z-variants. Other variants include glyphs with components rearranged (for example UTC00344, which rearranges the components of U+69AB but is pronounced the same and means the same), simplified versions of encoded characters (for example UTC00842), and ideographs which mean the same and are pronounced the same as encoded ideographs and have a sufficiently similar shape as to be easily mistaken for one another (for example UTC00399). This is a deliberately less strict, if somewhat more subjective, standard than is used for unification work. The Unicode field here indicates the encoded character of which this is a variant.

A status of W means that the ideograph is not suitable for encoding. An example here is UTC00118, which is used as a decoration in the novels *Xenocide* and *Children of the Mind* by Orson Scott Card. While the character does have an apparent intended meaning (something like "monster-killer"), it isn't suitable for encoding because of its ad hoc nature and lack of generalized use outside of the context of two specific English-language novels. The bulk of the characters with a status of W are Wenlin-specific Z-variants which should be represented (if at all), via a variation sequence defined by Wenlin, not by the UTC.

A status of X means the ideograph is a candidate for inclusion in an encoding proposal

post-dating Extension E.

2.2 The Source Field

The source field consists of source information, which consists of a source tag usually followed by a source-specific index string. Source tags and indices are separated by a space, and multiple source indices are separated by commas. Multiple sources are separated by asterisks.

The source tag may be a URI, in which case the index string is the date (year-month-day) when the URI was accessed. The source tag may also be a U-source index for cases where an ideograph was added to the U-source twice. The source tags beginning with a lowercase k correspond to fields within the Unihan database. Please consult [UAX38] for information on these sources and the format and meaning of the index strings.

The remaining sources are listed below. The left column contains the source tag. The center column contains bibliographic information for the source. The third column contains a description of source index, if any. The description frequently includes a regular expression which the index matches; see [UAX38] for more information.

Source Tag	Source Bibliographic Information	Source Index
ABC2	DeFrancis, John. *ABC Chinese-English Dictionary.* Honolulu: University of Hawai'i Press, 1999.	None
Adobe-Japan1	The Adobe-Japan1 glyph collection	The glyph index within the set
Cheng	Cheng Tso-Hsin, ed. *A complete checklist of species and subspecies of the Chinese birds.* Beijing: Science Press, 2000.	None
CN	Vũ Văn Kính, ed. *Đại Tự Điển Chữ Nôm.* Ho Chi Minh City: Nhà xuấ bản văn nghệ. 1998	A string matching the regular expression [01][0-9]{3}\.[0-9]{2} indicating the page and position on the page.
DYC	《說文解字·注》 Shuō Wén Jiě Zhì — Zhù [Annotated Qíng Dynasty recension of the Eastern Hàn Chinese analytic dictionary SWJZ]. 〖東漢〗許慎著 (121 AD)， 〖清〗段玉裁注 (1815)。 [上海古籍出版社, 1981.] See Cook (2003:461 ff; UMI #3105189) for complete references to the various editions: http://linguistics.berkeley.edu/~rscook/html/writing.html#EHC.	A string matching the regular expression [0-9]{3}\.[0-9]{2}[01] indicating the page and position on the page.
GB18030-2000	GB18030-2000	None
LDS	"Required Character List Supplied by The Church of Jesus Christ of Latter-day Saints"	The character index within the document
Shangwu	Huang Giangshang, ed. *Shangwu Xin Cidian.* Hong Kong: The Commercial Press, 1991. ISBN 962-07-0133-X	A string matching the regular expression [0-9]{3}\.[0-9]{2} indicating the page and position on the page.
TUS	The Unicode Consortium. *The Unicode Standard, Version 1.0, Volume 2.* Reading, Mass.: Addison-Wesley Publishing Company, 1992. ISBN 0-201-60845-6	The character's code point in the form U\+FA[0-9A-F]{2}
UDR	A defect report filed against the Unicode Standard or other direct communication with the Unicode editorial committee	None

WG2	A WG2 document	The document number
WL	Wenlin v. 3.1.8 http://www.wenlin.com	The PUA code point assigned the ideograph in the form E[0-9A-F]{3}
XHC	中国社会科学院语言研究所词典编辑室, ed.*Xiandai Hanyu Cidian.* Beijing: The Commercial Press. 2003	A string matching the regular expression [01][0-9]{3}\.[0-9]{2} indicating the page and position on the page.

References

[Glyphs] Glyph Table
For the latest version, see:
http://www.unicode.org/reports/tr45/tr45-glyphs-1.pdf

[Data] Text Data
For the latest version, see:
http://www.unicode.org/reports/tr45/tr45-sourcedata-1.txt

[UAX38] Unicode Han Database (Unihan)
http://www.unicode.org/reports/tr38/

Modifications

This section indicates the changes introduced by each revision.

Revision 1

- First version