Proposed Update Unicode Technical Report #45

U-SOURCE IDEOGRAPHS

Editor	John Jenkins 井作恆 (jenkins@apple.com)
Date	2011-07-25
This Version	http://www.unicode.org/reports/tr45/tr45-4.html
Previous Version	http://www.unicode.org/reports/tr45/tr45-3.html
Latest Version	http://www.unicode.org/reports/tr45/
Latest Proposed Update	http://www.unicode.org/reports/tr45/proposed.html
Tracking Number	4

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*This is a **draft** document which may be updated, replaced, or superseded by other documents at any time.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.

This document serves two purposes. First, it provides a formal reference to U-source ideographs, so that they may be referred to in other documents by their U-source identifiers. Second, it provides a public record of all ideographs which have been submitted to the Unicode Technical Committee for consideration. As such, it provides data on the nature, content, and disposition of these submissions.

The U-source database consists of three classes of CJK ideograph:

1. Ideographs which have been submitted to the UTC as potential candidates for encoding. Note that not all such ideographs are actually suitable for encoding. Those that are not have a status of "W".
2. Placeholder ideographs required to maintain continuity of U-source indices. Early versions of the U-source database allowed for the possibility of ideographs being withdrawn, generally because they had been added erroneously. Replacement ideographs were added in their place to keep any U-source index from being skipped. All such ideographs have a status of "W". (Ideographs are no longer withdrawn from the U-source database after they have been added.)
3. Placeholder ideographs required to provide encoded CJK Unified Ideographs with IRG source information. All CJK Unified Ideographs in ISO/IEC10646 are required to have at least one source identifier. Changes to IRG source information, however, can leave a given ideograph without any such sources. In such cases, the ideograph is included in the U-source database to guarantee it has at least one source. Such ideographs are indicated by a source prefix of "UCI" instead of "UTC".

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

- [Glyphs], a PDF showing the glyphs for the U-source ideographs. This document is a simple matrix with the representative glyph for a U-source ideograph and its identifier in each cell. The representative glyphs used are drawn in a modern style, such as is used by the IRG in its work. The use of modern forms for some characters originally drawn in a seal style should not be taken as implying any mechanism for the inclusion of seal forms as a whole in the Unicode Standard.
- [Data], a text file containing information regarding the ideographs. A detailed description of this file follows.

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" or "UCI", followed by a hyphen and five decimal digits, starting with 00001. Identifier numbers are not skipped, and are not reused. Identifier numbers are assigned sequentially. Ideographs whose prefix is "UTC" are either those submitted to the UTC for consideration or those included in the U-source database for placeholder purposes. Ideographs included to guarantee an IRG source reference have the prefix "UCI".
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, D, U, or V. The meaning of this field in these four 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, N, 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. The Unicode field here indicates the character's code point.

A status of D means that the ideograph is found in Extension D. The Unicode field here indicates the character's code point.

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 N means that the character is earmarked to be included to the IRG as part of the UTC's proposal for a future extension.

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 UTC-00344, which rearranges the components of U+69AB but is pronounced the same and means the same), simplified versions of encoded characters (for example UTC-00842), 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 UTC-00399). 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 UTC-00118, 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. Another example would be UTC-00643, which is a transcription error for U+5709.

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 final disposition of this ideograph has not been determined.

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.

Note that the sources listed here may not provide adequate evidence of use for IRG work. This is partly because characters listed here may not be suitable candidates for encoding, but also because IRG requirements for evidence have become increasingly stringent over time. Many of the characters in each of the sets encoded prior to Extension D do not have adequate evidence of use by current IRG standards.

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 xuất 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. Characters from the DYC were added to the U-source database as part of a preliminary exploration of the possibility of encoding them. They will not be used for any effort to actually encode the contents of the DYC and should not be taken as the basis for any such encoding.	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,*	The character's code point in the form U\+FA[0-9A-F]{2}

	Volume 2. Reading, Mass.: Addison-Wesley Publishing Company, 1992. ISBN 0-201-60845-6	
UDR	A defect report filed against the Unicode Standard or other direct communication with the Unicode editorial committee	None
XHC	《现代汉语词典》 [Xiàndài Hànyǔ Cídiǎn = XHC; 'Modern Chinese Dictionary']. 中国社会科学院语言研究所词典编辑室编 [Chinese Academy of Social Sciences, Linguisitics Research Institute, Dictionary Editorial Office, eds.]. 北京: 商务印书馆, 2002. This is a later edition of the kXHC1983 source.	The page and position information in the format used by the kXHC1983 source
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}

References

[Data] Text Data
For the latest version, see:
http://www.unicode.org/reports/tr45/tr45-sourcedata-4.txt

[Feedback] Reporting Form
http://www.unicode.org/reporting.html
For reporting errors and requesting information online.

[Glyphs] Glyph Table
For the latest version, see:
http://www.unicode.org/reports/tr45/tr45-glyphs-4.pdf

[Reports] Unicode Technical Reports
http://www.unicode.org/reports/
For information on the status and development process for technical reports, and for a list of technical reports.

[UAX38] UAX #38: *Unicode Han Database (Unihan)*
http://www.unicode.org/reports/tr38/

[Unicode] The Unicode Standard
For the latest version, see:
http://www.unicode.org/versions/latest/
For the 5.2.0 version, see:
http://www.unicode.org/versions/Unicode5.2.0/

[Versions] Versions of the Unicode Standard
http://www.unicode.org/versions/
For information on version numbering, and citing and referencing the Unicode Standard, the Unicode Character Database, and Unicode Technical Reports.

Acknowledgements

John Jenkins is the author of the initial version and has added to and maintains the text of this report.

Modifications

The following summarizes modifications from the previous revision of this document.

Revision 4:

- Inclusion of characters with an index prefix of "UCI".
- Clarified the use of dummy characters as placeholders.
- General updates to the data files

Revision 3:

- Changes in character status per actions taken at WG2 meeting 54
- Clarified nature of characters from the DYC
- Clarified relationship between UTC sources and IRG evidence

Revision 2:

- First approved version.
- Changes in character status per actions taken at IRG meeting 31.
- Revisions per input from UTC.

Revision 1:

- First draft version.
