Technical Reports

L2/19-357

Proposed Update Unicode® Standard Annex #45

U-SOURCE IDEOGRAPHS

Version	Unicode 13.0.0 (draft 1)
Editor	John H. Jenkins 井作恆
Date	2019-05-16
This Version	http://www.unicode.org/reports/tr45/tr45-22.html
Previous Version	http://www.unicode.org/reports/tr45/tr45-21.html
Latest Version	http://www.unicode.org/reports/tr45/
Latest Proposed Update	http://www.unicode.org/reports/tr45/proposed.html
Revision	22

Summary

This annex describes UTC-source ideographs as used by the Ideographic Research 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*This is a **draft** document which may be updated, replaced, or superseded by other documents at any time.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 Standard Annex (UAX)** forms an integral part of the Unicode Standard, but is published online as a separate document. The Unicode Standard may require conformance to normative content in a Unicode Standard Annex, if so specified in the Conformance chapter of that version of the Unicode Standard. The version number of a UAX document corresponds to the version of the Unicode Standard of which it forms a part.*

Please submit corrigenda and other comments with the online reporting form [Feedback]. Related information that is useful in understanding this annex is found in Unicode Standard Annex #41, "Common References for Unicode Standard Annexes." 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]. For any errata which may apply to this annex, see [Errata].

Contents

1 Introduction

This annex describes a subset of IRG U-source ideographs as used by the Ideographic Research 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. Each time the IRG begins the process of preparing a new CJK Unified Ideographs extension, IRG members submit a set of characters for potential inclusion in that extension. The IRG classifies these into sources, one for each submitter, e.g., the J-source for Japan, the V-source for Vietnam.

The IRG U-source currently includes submissions from two organizations: the Unicode Technical Committee and the SAT Daizōkyō Text Database Committee. The organization which submitted such characters can be determined by the prefix of the character's U-source identifier. If the identifier begins with "USAT", the character was submitted by the SAT Daizōkyō Text Database Committee; otherwise, it was submitted by the Unicode Technical Committee. For example, U+20991 has the U-source identifier "USAT-00061" and so was submitted by the SAT Daizōkyō Text Database Committee. U+FA2D has the U-source identifier "UTC-00936" and so was submitted by the UTC.

The IRG U-source also includes characters submitted by neither organization. All these characters are encoded because they were originally submitted to the IRG by some other body. Some of these are characters which were submitted *to* the UTC for consideration but were not submitted *by* the UTC to the IRG. (The IRG refers to such cases as "horizontal extensions.") Others were left without a formal IRG source by changes made by the IRG in its source-mappings; these were "adopted" by the UTC as explained below.

Formally, the IRG U-source may be considered to consist of two subsources: the SAT- or USAT-source, which includes those characters submitted by the SAT Daizōkyō Text Database, and the UTC-source, which includes everything else. Note that there is no requirement that the U-source include *all* USAT- or UTC-source ideographs. As is generally the case, the IRG only takes cognizance of those portions of these subsources brought to its attention.

Identifier Prefix	Source	Responsible Body
UTC	UTC-source	Unicode Technical Committee
UCI	UTC-source	Ideographic Research Group
UK	UTC-source	Andrew West (formal affiliation?)

USAT	USAT-source	SAT Daizōkyō Text Database

This document serves two purposes. First, it provides a formal reference to UTC-source ideographs, so that they may be referred to in other documents by their UTC-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 UTC-source database consists of four 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 as a CJK Unified Ideograph. Those that are not have a status of "W".
2. Placeholder ideographs required to maintain continuity of UTC-source indices. Early versions of the UTC-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 UTC-source index from being skipped. All such ideographs have a status of "W". (Ideographs are no longer withdrawn from the UTC-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 UTC-source database to guarantee it has at least one source. Such ideographs are indicated by a source prefix of "UCI" instead of "UTC". Later changes to IRG source information may restore non-UTC-source sources or provide new ones for such ideographs. In such cases, the ideograph retains the "UCI" prefix, but the encoded character loses its kIRG_USource reference in the Unihan database and ISO/IEC 10646.
4. Ideographs submitted by the UK to the IRG for Working Set 2015. These ideographs are included to provide a standard, stable source reference for IRG purposes. Such ideographs are indicated by a source prefix of "UK" instead of "UTC".

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

- [Glyphs45], a PDF showing the glyphs for the UTC-source ideographs. This document is a simple matrix with the representative glyph for a UTC-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 [Unicode].
- [RSChart45], a PDF providing a set of radical-stroke charts for the UTC-source ideographs. These charts use the glyphs used in the main glyph chart.
- [Data45], 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 eight fields separated by semicolons.

1. The ideograph's UTC-source identifier. This consists of the letters "UTC", "UK", 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 UTC-source database for placeholder purposes. Ideographs included to guarantee an IRG source reference have the prefix "UCI".

2. A string indicating the ideograph's current status. These are described below.
3. A Unicode code point. This field is generally empty. Its interpretation depends on the character's status and is documented 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.
8. General comments regarding the ideograph.

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 A, B, C, Comp, D, E, F, G, N, U, V, W, X, ~~UNC-2013, UNC-2015,~~ UK-2015, WS-2017, and strings matching the regular expressions "UTC-\d{5}", "UCI-\d{5}", and UK-\d{5}; new values may be added in the future.

Status	Meaning	Value of Unicode Field
A	Encoded in Extension A	The character's code point
B	Encoded in Extension B	The character's code point
C	Encoded in Extension C	The character's code point
Comp	Encoded as a compatibility ideograph	The character's code point
D	Encoded in Extension D	The character's code point
E	Encoded in Extension E	The character's code point
F	Encoded in Extension F	The character's code point
G	Submitted by the UTC for IRG Working Set 2015	The code point of a character to which this is related, generally as a variant
N	Earmarked to be included in a proposal from the UTC to the IRG for a future extension	The code point of a character to which this is related, generally as a variant
U	Encoded in the URO	The character's code

		point
V	A variant of an encoded ideograph (see below)	The code point of the character of which this is a variant
W	Not suitable for encoding as a CJK Unified Ideograph (see below)	The code point of a character to which this is related, generally as a variant
X	Appropriate disposition has not been determined	The code point of a character to which this is related, generally as a variant
~~UNC-2013~~	~~Included in the UTC's 2013 "Urgently Needed Characters" proposal to the IRG~~	~~The character's code point~~
~~UNC-2015~~	~~Included in the UTC's 2015 "Urgently Needed Characters" proposal to the IRG~~	~~The code point of a character to which this is related, generally as a variant~~
UK-2015	Submitted by the UK for IRG Working Set 2015	The code point of a character to which this is related, generally as a variant
WS-2017	Submitted by the UTC for IRG Working Set 2017	The code point of a character to which this is related, generally as a variant
Strings matching the regular expressions "UTC-\d{5}", "UCI-\d{5}", and "UK-\d{5}".	Duplicate entries deprecated in favor of other entries; the status value is the identifier of the non-deprecated character	The character's code point, or the code point of a character to which this is related, generally as a variant

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.

A status of W means that the ideograph is not suitable for encoding as a CJK Unified Ideograph. An example here is UTC-00326, which is a nonce form specifically coined for use with Figure 18-9, *Using the Ideographic Description Characters*, in [Unicode] and to fill an empty slot in the UTC-source database. While the character does have an intended meaning ("frog at the bottom of a well"), it isn't suitable for encoding because of its *ad hoc* nature and lack of generalized use. Another example would be UTC-00643, which is a transcription error for U+5709. ~~Characters such as UTC-03160 through UTC-03166 are suitable for encoding, but not as CJK Unified Ideographs. Because they occur only in the context of Gongche music notation, they are best handled by encoding a full set of Gongche notation characters in a separate block.~~ 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.

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 UTC-source index for cases where an ideograph was added to the UTC-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-CNS1	The Adobe-CNS1 glyph collection	The glyph index within the set matching the regular expression (C\+)?[0-9]{1,5}

Adobe-Japan1	The Adobe-Japan1 glyph collection	The glyph index within the set matching the regular expression (C\+)?[0-9]{1,5}
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 Điên Chữ Nôm.* Ho Chi Minh City: Nhà xuâ 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 UTC-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	[Unicode]	The character's code point matching the regular expression U\+2?[0-9A-F]{4}
UDR	A defect report filed against the Unicode Standard or other direct communication with the Unicode editorial committee	None
UTCDoc	A UTC document	The document number optionally followed by a decimal index for the character within the document
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, Linguistics 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 matching the regular expression E[0-9A-F]{3}

2.3 The Comments Field

The comments field is a general-purpose, unstructured field. It is generally empty. It can contain any Unicode character other than tabs, semicolons, and any line-break character. The purpose of this field is to provide any additional relevant information for an ideograph which is not included in any other fields. For example:

1. The comment field for UTC-00119 indicates why it is inappropriate for encoding
2. The comment field for UTC-00595 clarifies its variation relationship to other characters
3. The comment field for UTC-02981 contains a kMandarin value pending encoding

References

For references for this annex, see Unicode Standard Annex #41, "Common References for Unicode Standard Annexes."

Acknowledgements

The UTC gratefully acknowledges the contributions of Eiso Chan, Henry Chan, Jaemin Chung, Lee Collins, Richard Cook, Jing Zuoheng, Ken Lunde, Ming Fan, William Nelson, Andrew West, and others to the UTC-source database.

Modifications

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

Revision 22

- **Proposed update** for Unicode 13.0.0.
- Added table summarizing U-source identifiers
- Added 'Comp' status value

- Removed UNC-2013 and UNC-2015 status values
- Added description of radical-stroke charts

Revision 21

- **Reissued** for Unicode 12.0.0.
- Added values A and B to the status field.
- Added documentation of the new comments field added to the data file.

Revision 20 being a Proposed Update, only changes between Revisions 19 and 21 are listed here.

Modifications for previous versions are listed in those respective versions.
