



Proposed Update Unicode® Standard Annex #45

U-SOURCE IDEOGRAPHS

Version	Unicode 10.0.0
Editor	John H. Jenkins 井作恆 (jenkins@apple.com)
Date	2017-02-06
This Version	http://www.unicode.org/reports/tr45/tr45-16.html
Previous Version	http://www.unicode.org/reports/tr45/tr45-15.html
Latest Version	http://www.unicode.org/reports/tr45/
Latest Proposed Update	http://www.unicode.org/reports/tr45/proposed.html
Revision	<u>16</u>

Summary

This annex describes **UTC**-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*This is a **draft** document which may be updated, replaced, or superseded by other documents at any time.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 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](#)
 - 2 [Text File Data](#)
 - 2.1 [The Status Field](#)
 - 2.2 [The Source Field](#)
- [References](#)
[Acknowledgements](#)
[Modifications](#)

1 Introduction

This annex describes [a subset of IRG](#) 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. 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.

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 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 [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.)

- 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".

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 the Unicode Standard.
- [[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 seven fields separated by semicolons.

- The ideograph's **UTC**-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 **UTC**-source database for placeholder purposes. Ideographs included to guarantee an IRG source reference have the prefix "UCI".
- A string indicating the ideograph's current status. These are described below.
- A Unicode code point. This field is **generally empty. Its interpretation depends on the character's status and is documented below.**
- A radical-stroke index for the ideograph, as described in [[UAX38](#)].
- A KangXi dictionary index for the ideograph, as described in [[UAX38](#)].
- An ideographic description sequence (IDS) for the ideograph, if one can be generated.
- 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, F, **G**, N, U, V, W, X, UNC-2013, UNC-2015, UK-2015, and strings matching the regular expressions "UTC-\d{5}" and "UCI-\d{5}"; new values may be added in the future.

Status	Meaning	Value of Unicode Field
C	Found in Extension C	The character's code point
D	Found in Extension D	The character's code point
E	Found in Extension E	The character's code point

F	Found 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	Otherwise encoded in Unicode	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 (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
Strings matching the regular	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

expressions "UTC-\d{5}" and "UCI-\d{5}"	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. 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.

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. <i>ABC Chinese-English Dictionary</i> . 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. <i>A complete checklist of species and subspecies of the Chinese birds</i> . Beijing: Science Press, 2000.	None
CN	Vũ Văn Kính, ed. <i>Đại Tự Điển Chữ Nôm</i> . 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 UTC –source database as part of a preliminary exploration of the possibility of encoding them. They will not be used for any effort to	A string matching the regular expression [0–9]{3}\.[0–9]{2}[01] indicating the page and

	actually encode the contents of the DYC and should not be taken as the basis for any such encoding.	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. <i>Shangwu Xin Cidian</i> . Hong Kong: The Commercial Press, 1991. ISBN 962–07–0133–X	A string matching the regular expression <code>[0–9]{3}\.[0–9]{2}</code> indicating the page and position on the page.
TUS	[Unicode]	The character's code point matching the regular expression <code>U\+2?[0–9A–F]{4}</code>
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}

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 Henry Chan, Lee Collins, Richard Cook, Jing Zuoheng, Ken Lunde, Ming Fan, Andrew West and others to the UTC-source database.

Modifications

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

Revision 16

- Updated terminology to reflect the difference between the IRG's U-source and the set of characters submitted by the UTC (now referred to as the "UTC-source").
- Updates to contents and to status values reflecting decisions of the UTC and IRG.
- Proposed update** for Unicode 10.0.0.

Revision 15

- **Reissued** for Unicode 9.0.0.
- Additional values for status field.

Revision 14 being a Proposed Update, only changes between Revisions 13 and 15 are listed here.

Revision 13

- **Reissued** for Unicode 8.0.0.
- Revision and correction of table styles.

Revision 12

- **Reissued** for Unicode 7.0.0.
- Clarified meaning of status field.

Revision 11 being a Proposed Update, only changes between Revisions 10 and 12 are listed here.

Revision 10

- **Reissued** for Unicode 6.3.0.
- Changed status field from a single character to a string.
- Added new status value of "UNC-2013".
- Added UTCDoc source.

Revision 9 being a Proposed Update, only changes between Revisions 8 and 10 are listed here.

Revision 8

- **Reissued** for Unicode 6.2.0.
- Changed status from UTR to UAX.
- Added a clear definition for U-Source.
- Added a new category for Extension F.

Revision 7 being a Proposed Update, only changes between Revisions 6 and 8 are listed here.

Revision 6

- Corrected Adobe-CNS1 and Adobe-Japan1 source references.
- Updated data file and glyph chart to reflect the results of IRG meeting #37.
- Corrected reference to the Unicode Standard.

Revision 5

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

Revision 4 being a Proposed Update, only changes between Revisions 3 and 5 are listed here.

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.

Copyright © 2017  Unicode, Inc. All Rights Reserved. The Unicode Consortium makes no expressed or implied warranty of any kind, and assumes no liability for errors or omissions. No liability is assumed for incidental and consequential damages in connection with or arising out of the use of the information or programs contained or accompanying this technical report. The Unicode [Terms of Use](#) apply.

Unicode and the Unicode logo are trademarks of Unicode, Inc., and are registered in some jurisdictions.