GlyphWiki and OpenType: A Collaborative Glyph Development Environment and its Font Exporting System

Taichi KAWABATA, NTT Network Innovations Laboratories <Kawabata.taichi@lab.ntt.co.jp>
Koichi KAMICHI, Daito-Bunka University <kamichi@fonts.jp>

Abstract
Today, approximately 75,000 CJK ideographs have been encoded in ISO/IEC 10646 and Unicode. However, there are still many more unencoded ideographs, many of which will remain unencoded because they are considered variants of encoded ideographs, and there is still some demand to distinguish subtle minor design differences. GlyphWiki, developed by Koichi Kamichi of Daito-Bunka University, is a wiki-based collaborative glyph design system that meets such demands. By using this system, GlyphWiki published world’s first fonts that cover all CJK ideographs in UCS/Unicode. GlyphWiki enables its users to design their own character, name it, and publish it as an OpenType font or SVG file. This paper describes the architecture of GlyphWiki, its glyph-naming system and how the OpenType features of published font correspond to these Wiki-based naming schemes.

Keywords
OpenType, GlyphWiki, Wiki, Ideographs, HTML5, CSS3, Text Encoding

Introduction
Recently, electronic books have become widely popular throughout the world. In 2011, EPUB3, an electronic book standard, was published, and it incorporates various East Asian text features by adopting HTML5 and related standards, and various EPUB3-based EBook readers have been released to the market. Demand for encoding not only modern, technical documents, but also classical, literary documents has become high. However, concerning East Asian text involving ideographic characters, there is always the problem of handling unencoded ideographs.

CJK Unified Ideographs
The first international ideographic character set was standardized as Unicode or ISO/IEC 10646 in 1992. These characters are called CJK Unified Ideographs, and included 20,902 ideographs. These ideographs were collected from various East Asian national and regional standards. They are guaranteed to be interoperable with these standards (round-trip conversion), and similar characters within these standards are unified (Han Unification). Since then, CJK Unified Ideographs have been continuously expanded. In 1998, Extension A standardized an additional 6,582 ideographs. In 2000, Extension B standardized 42,711 more ideographs. In 2006, Extension C standardized 4,149 more ideographs. In 2009, Extension D added 222 urgently needed ideographs. Extension E is expected to be standardized in 2013, and it will include more than 5,000 ideographs. In total, approximately 75,000 ideographs are already encoded. Furthermore, work on Extension F is likely to include more than 10,000 ideographs. These new ideographs will mainly come from regional or personal names that are used in government databases, or ideographs contained in dictionaries.

During the standardization process, unification criteria were also developed, to more clearly determine whether some specific glyphs should be unified or not.
Unencoded ideographs will never be exhausted
Although a large amount of ideographs are encoded, unencoded ideographs will never be exhausted. Classical texts sometimes contain ideographs that are not encoded, or modern authors may coin new ideographs in their works. We must establish a proper way to handle unencoded ideographs when encoding East Asian text, such as to create and design new glyphs, ways to share them with others, and ways to use them in encoded text in a future-sustainable way. That’s why GlyphWiki was developed.

Unencoded Ideographs

Unencoded ideographs may appear anywhere regardless of modern or classical texts. Because of this, digitization of East Asian text sometimes involves not only encoding of text data, but also creating unencoded glyphs as a font resource.

Unencoded characters may be classified into three categories. One is unencoded, but a unifiable ideographic variant according to the unification criteria of the ISO standard (ISO/IEC 10646 Annex S). Another is also a variant of encoded ideographs, but not unifiable. The third one includes ideographs that are totally new, and are not variants of existing ideographs.

Several solutions may be considered to describe such unencoded ideographs in digitized text:

1. Embedded Glyph Images
   Unencoded glyphs can be handled as an embedded image file. This solution is simple, but images usually do not behave as characters when they are laid out on pages. Also, and perhaps more importantly, they cannot be searched nor copied as text.

2. Create User-Defined Characters (UDC) encoded in the Private Use Area (PUA).
   This is also a common solution. However, when this solution is applied, the font and text must be tied together to preserve the integrity of the text. Tools that extract text and separate it from the accompanied font, such as search engines, cannot handle such ideographs. In the future, if such ideographs are encoded, a publisher may have to check old documents with human eyes to determine whether they can be replaced with newly-encoded ones. When multiple documents considered, PUA code points may collide, resulting in different or unexpected glyphs.

3. Ideographic Description Sequence (IDS)
   Ideographic Description Sequences, comprised of Ideographic Description Characters (IDC) and Description Characters (DC), have the ability to describe unencoded ideographs in a human-readable way. Its mechanism is simple. Usually, ideographs are made up of smaller components, aligned vertically or horizontally. IDCs describe the structure of ideographs, and DC represents a component of ideographs. For example, “横” can be described as “木 竹”, and “道” can be described as “辵 道”. An IDS is useful as a structural representation of an unencoded ideograph, and easily searchable and replaceable, but to display a complete glyph that corresponds to an IDS requires special mechanisms.

4. Ideographic Variation Sequence (IVS)
   An Ideographic Variation Sequence is a standardized method for encoding glyphs that is described in Unicode Technical Standard #37. An IVS is comprised of a CJK Unified Ideograph followed by a Variation Selector (VS). To use an IVS, it must be registered in a collection within the Ideographic Variation Database (IVD), which is maintained by the Unicode Consortium. IVSes are arranged according to IVD collections. As of 2012, two IVD collections have been registered. One is the Adobe-Japan1 collection, registered by Adobe, and the other is the Hanyo-Denshi collection, registered by the Hanyo Denshi (汎用電子) Committee of Japan. If the desired ideograph is already registered in the IVD, then one may use its IVS to encode it. An IVS is stable, can be copied as “plain text”, and can be searched.
It should be noted that standardization activities concerning CJK Unified Ideographs are still ongoing, and there is always a chance that unencoded ideographs appearing in legacy documents may be encoded in a future extension.

**GlyphWiki**

GlyphWiki is a collaborative glyph-design system developed by Kamichi Koichi of Daito Bunka University. Its user interface is similar to that of existing Wiki-based systems, but its main purpose is to let users design glyphs freely, share them with other users, and let other documents reference or embed them.

---

**Figure 1** Glyph Editing Screen
Today, GlyphWiki contains nearly 300,000 glyphs designed by various people. Several published books, including ones on Chinese medicines, have adopted glyphs developed by this system.

Once glyphs are designed, they can be exported as a single- or multiple-glyph TrueType font, SVG font, SVG image, PNG image or JSON representation that can be used by various Web Services.

**GlyphWiki Glyph Naming Rules & Conventions**

With GlyphWiki, users can create glyphs for any ideographs. These glyphs can be classified into two categories: protected (private) glyphs and public glyphs. Protected (private) glyphs are ones whose name follows the form `username-glyphname`. Protected glyphs can be referenced by anyone, but only the original user can edit them. Thus, private glyphs are protected from modification by other users. Public glyphs are ones whose names do not begin with `username`. People can name their own glyph freely, regardless of whether they are public or private. Public glyphs are often taken from existing regional standards or established dictionaries. In this case, their names follow some rules and conventions, determined by discussions among members. For example, glyphs for ideographs from the `zihai-XXXXYY` dictionary have the name `zihai-XXXXYY`, where `XXX` is the page number and `YYY` is the entry number within the specified page. Similarly, `lx-XXXXYY` is for 康熙字典 (KangXi Dictionary), `koseki-XXXXXX` is for 戶籍統一文字, and so on.

Unicode and related glyphs adopt a complex naming scheme. In GlyphWiki, Unicode ideographs have the name `uXXXX`, where `XXX` is the hexadecimal code point. For example, glyph with the name `u4e00` denotes “一”-™. If a single glyph is comprised of multiple Unicode code points, such code points are connected with a hyphen. This includes IDSes, IVSes and ligatures. For example, `u5050-ue0102` denotes the IVS of the glyphs for “些”, whose sequence identifier is “FT2069” according to IVD.

Unicode glyph names may contain a wide variety of suffixes, in a specified order. In ISO/IEC 10646 standard, CJK Unified Ideographs are arranged in multi-column form according to region. When a region-specific glyph is designed, its name includes a suffix letter that represents its region. For example, `uXXXX-g` is a Chinese-style glyph of `uXXXX`. Similarly, `uXXXX-t` is for Taiwan, and `uXXXX-` is for Japan and so on.

Glyphs that behave as components include a suffix with two decimal numbers that indicate the component type. For example, `uXXXX-01` is the left-side component of `uXXXX`. The following table shows the component indicator naming rules:

<table>
<thead>
<tr>
<th>Indicator</th>
<th>Meaning</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>-01</td>
<td>Left component</td>
<td>u91d1-01: ሱ́</td>
</tr>
<tr>
<td>-02</td>
<td>Right component</td>
<td>u200a2-02:</td>
</tr>
<tr>
<td>-03</td>
<td>Upper component</td>
<td>u3ac1-03:</td>
</tr>
<tr>
<td>-04</td>
<td>Lower component</td>
<td>u3687-04:</td>
</tr>
<tr>
<td>-05</td>
<td>Outside component</td>
<td>u3ac5-05:</td>
</tr>
<tr>
<td>-06</td>
<td>Inside component</td>
<td>u4ea4-06:</td>
</tr>
<tr>
<td>-07</td>
<td>General component</td>
<td>u4ee4-07:</td>
</tr>
<tr>
<td>-08</td>
<td>Vertically extended component</td>
<td>u5355-08:</td>
</tr>
<tr>
<td>-09</td>
<td>Horizontally extended component</td>
<td>u535d-09:</td>
</tr>
<tr>
<td>-10,11,14,15</td>
<td>Miscellaneous</td>
<td></td>
</tr>
</tbody>
</table>

Table 1 GlyphWiki component indicators

Unifiable variants have a suffix of `-var-XXX` where `XXX` is a decimal number. Non-unifiable variants have a suffix of `-itaiji-XXX` ("itaiji" means variants in Japanese). Vertical forms have a suffix of `-vert`. These rules are summarized in BNF forms:
<glyph-name>::=<base-char>[suffix>
<base-char>::=<single-char>|<IDS>|<IVS>|<ligature>
<single-char>::="uXXXX"(XXXX is hexadecimal)
<IDS>::=<IDC>{"."<single-char>}+
<IDC>::="u2ff0"|"u2ff1"|...|"u2ffb"(Ideographic Description Character)
<IVS>::=<single-char>"-ue01XX"(XX is hexadecimal)
<ligature>::=<single-char>{"."<single-char>}+
<suffix>::=["."[<region>]<[component]>][<unifiable>|<non-unifiable>][<vert>]
<br><region>::="g"|"t"|"j"|"k"|"v"|"h"
<br><component>::="01"|"02"|...|"11"|"14"|"15"
<br><unifiable>::="-var-XXX"(XXX is decimal)
<br><non-unifiable>::="-itaiji-XXX"(XXX is decimal)
<br><vert>::="-vert"

For example, the glyph name ‘u4ea1-g01-var-001’ means the first unifiable variant of the Chinese-style left-side component representation of U+4EA1. Some glyphs include extensive variations registered by GlyphWiki users.

Hanazono Mincho Font
Hanazono Mincho is a font produced from GlyphWiki’s collaborative glyphs. Currently, it is the only font that fully supports all CJK Unified Ideographs and IVS glyphs. A revision of this font to support the GSUB features as described below is planned.

OpenType and CSS3 Fonts
OpenType and CSS3 Fonts are key technologies for modern digital publications. They are adopted in EPUB3 and other digital publication standards, and are also important for digitization of East Asian text. Their flexible ability to describe a wide variety of text easily enables the textual representation of unencoded characters.

OpenType
OpenType, which was co-developed by Adobe and Microsoft, has become the preferred font format, and has been adopted by a wide variety of operating systems and applications. One of its most important characteristics is its customizable glyph substitution (GSUB) and glyph positioning (GPOS) features, enabling ligatures and a wide variety of variant glyphs.

GSUB features consist of multiple lookup tables, which convert glyphs to other glyphs. Lookup tables can be classified into five types: “single”, “multiple”, “alternate”, “ligature” and “contextual”. The “single”-type lookup table converts one glyph into another. The “multiple”-type lookup table converts one glyph into multiple glyphs. The “alternate”-type lookup table converts one glyph into one of several glyphs, at the user’s or application’s discretion. The “ligature”-type lookup table converts multiple glyphs to a single glyph. The “contextual”-type lookup table converts one or more glyphs according to a specified context.

Each GSUB feature is registered as a four-character tag. For example, the ‘vert’ feature converts standard (horizontal) glyphs to the vertical forms. Whether specific OpenType features should be turned on by default can be determined by the application. The order of GSUB features (and their lookup tables) is determined by each OpenType font.

Feature tags are registered and appear in the OpenType Specification. The followings are some examples of OpenType features that will be mentioned later in this paper.
Among the various features, the 'locl' feature is peculiar. It can convert “default” glyphs to “localized” glyphs. Originally, lookup tables are associated with specific languages. When multiple lookup tables of 'single' type with different languages are included in the 'locl' feature, default (encoded) glyphs can be converted to 'localized' glyphs in accordance with the specified languages.

OpenType also supports Ideographic Variation Sequences (IVS). IVS support in OpenType is done through the use of the Format 14 'cmap' subtable, not with GSUB features.

### CSS3 Fonts (CSS Fonts Module Level 3)

CSS3 Fonts is a key technology of HTML5 that describes how fonts should be specified in HTML. It defines a wide variety of font properties, from very basic font selection and a font-matching algorithm to dynamic font loading. Besides these various properties, it includes in its specification the ability to control OpenType GSUB features. This is called the 'font-feature-settings' property, which can directly specify which GSUB features should be activated.

For example, to activate the 'trad' GSUB feature and to invoke its third alternative and also activate the 'locl' feature for Japan for font ‘foo’, one can specify this as follows in HTML5:

```html
<span xml:lang="ja_JP" style="font-family: foo; font-feature-settings: 'trad' 3, 'locl'>漢</span>
```

This CSS3 property enables the invoking various glyphs while preserving text consistency, and is quite suitable to describe East Asian ideograph variants.

Currently, CSS3 Fonts is still a public working draft, and as of this writing, only Google Chrome Canary partially supports the 'font-feature-settings' property. However, it is expected that in the near future, after CSS3 Fonts is stabilized, many web browsers will support this feature.

### GlyphWiki and OpenType

Originally, GlyphWiki has provided simple TrueType font production facilities for its own fonts. However, it has defects that are common to PUA-based solutions, such as not being searchable, not combinable with other text, and not future-sustainable.

To cope with this problem, we’ve designed a new font publishing mechanism that provides the following solutions:

1. HTML5- and CSS3-compliant, so that it can cope with modern browsers and future digital publications.
2. Glyphs can be uniquely specified in HTML/CSS, so that they can be safely copied from one to another.
3. It should support all properly named Unicode-related GlyphWiki glyphs.
4. Fonts are extensible and scalable. GlyphWiki should produce minimal fonts needed for some specific document.

To meet these demands, we’ve came up with the idea to map GlyphWiki names to corresponding OpenType features. The following table shows how GlyphWiki Unicode glyph names will be mapped to equivalent OpenType features. We’ve selected following GSUB features for our mappings.

<table>
<thead>
<tr>
<th>GlyphWiki name</th>
<th>OpenType table/feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>uXXXX-ue010.XX</td>
<td>cmap (Format 14)</td>
<td>Ideographic Variation Sequence will be mapped to the Format 14 ‘cmap’ subtable.</td>
</tr>
<tr>
<td>u2ffX-uXXXX-uYYYY....</td>
<td>ccmp</td>
<td>Ideographic Description Sequence will be mapped to the OpenType ‘ccmp’ feature.</td>
</tr>
<tr>
<td>uXXXX-uYYYY</td>
<td>liga</td>
<td>Ligatures will be mapped to the OpenType ‘liga’ feature.</td>
</tr>
<tr>
<td>uXXXX-g, etc.</td>
<td>locl</td>
<td>Regional glyphs will be mapped to OpenType ‘locl’ feature.</td>
</tr>
<tr>
<td>uXXXX-01, uXXXX-02, etc..</td>
<td>ss01-ss15</td>
<td>Component indicator will be mapped to the OpenType ‘ssXX’ features.</td>
</tr>
<tr>
<td>uXXXX-var-XXX</td>
<td>salt</td>
<td>Unifiable variants will be mapped to the OpenType ‘salt’ feature.</td>
</tr>
<tr>
<td>uXXXX-itaiji-XXX</td>
<td>trad</td>
<td>Non-unifiable variants will be mapped to the OpenType “trad” feature.</td>
</tr>
<tr>
<td>uXXXX-vert</td>
<td>vert</td>
<td>Vertical forms will be mapped to the OpenType ‘vert’ feature</td>
</tr>
</tbody>
</table>

Table 3 GlyphWiki names vs. OpenType GSUB features

These glyph name suffixes can be combined in a specified order (except for the base character), so that all features may be used (e.g. u2ff0-u4e00-u4e01-g03-var-001-vert).

The OpenType ‘liga’ feature is suitable for combining characters when using GlyphWiki glyph names. The followings are example of how GlyphWiki ‘u3042-u3099’, ‘u3042-u309a’, etc. are converted:
The `ccmp' feature is a ligature feature for glyph composition, which is suitable for composing an IDS into a single glyph. With this feature, unencoded ideographs can be converted into a single glyph, while maintaining the text encoding. It also has the virtue of future-sustainability, that when new ideographs are encoded in a future version of the standard, they can be easily replaced by simple search-and-replace mechanisms. The following figure shows how GlyphWiki glyphs named ‘u2ff0-u2008-e-u53c8’ and ‘u2ff0-u200a4-u6bb3’ are realized as the OpenType ‘ccmp’ feature.

The `locl' feature is also suitable for the regional indicator of GlyphWiki names. For example, ‘uXXXX-j’ will be mapped to a lookup table tagged with Japan, and ‘uXXXX-g’ will be mapped to a lookup table tagged with China.

We’ve decided to map the component indicator of GlyphWiki glyph names to the ‘ss01’ to ‘ss15’ stylistic sets. The component indicator denotes behavior of glyphs as components or a specific style, and it is thus suitable to use the stylistic sets.

The suffix “-itaiji-XXX” denotes non-unifiable variants and provides more than one choice for substitution, so it should be mapped to a GSUB feature with an ‘alternate’-type lookup table. We’ve chosen the ‘trad’ feature for this suffix. The original purpose of the ‘trad’ feature is to denote ‘traditional’ glyphs, so it may not be suitable with its original purpose. However, because several OpenType fonts adopt this feature to convert glyphs to non-unifiable glyphs, it may not surprise the user that this feature is used for non-unifiable variants.
When the produced font is used with HTML, text represented with the `font-feature-settings` property can be copied from one document to another without worrying about losing glyph information.

**Font Production Workflow**

The production of fully-featured OpenType fonts involves complex procedures. We first adopted FontForge, and later AFDKO (Adobe Font Development Kit for OpenType), for GlyphWiki-based OpenType font production (we call it Hanazono Mincho). As Hanazono Mincho supports IVS, a CID-keyed font must be used. (AFDKO supports only CID-keyed fonts for production of IVS-enabled fonts.) The following figure shows the basic workflow to create an OpenType font from GlyphWiki data. First, GlyphWiki dumps KAGE (basic skeletal glyph structure data), which is processed by `dumpucs.pl` to identify duplicate glyphs (temp.alias), assigns temporary Unicode values (from Plane 15 PUA code points) for non-standard glyphs (temp.map), and assigns unique Unicode values to such glyphs and extracts appropriate skeletal data (temp.sources). The KAGE
engine `makeSVGFont.pl' converts the `temp.sources' KAGE data to the SVG font, and then it is converted to Compact Font Format (CFF) font data by using AFDKO’s `tx' tool. The `tx' tool also produces a “GIDs vs. temporary Unicode values” mapping table called `temp.dump'. By using `temp.alias', `temp.map' and `temp.dump', the tool `gw-afdko.el' parses GlyphWiki glyph names, maps them to GIDs (CIDs) produced by `tx’, and creates an appropriate GSUB “features” file, IVS data file and a UTF-32 CMap resource file. The generated CFF file (temp.cff) is processed by the AFDKO `checkOutlines' and `autohint' tools to remove intersections and to insert hinting information, respectively, and is then converted to a CID-keyed font by using AFDKO `mergefonts' tool, mapping all GIDs to corresponding CIDs. All of these produced files are used as input to the resource AFDKO `makeotf' tool to create a fully-functional OpenType font.

![GlyphWiki font production workflow](image)

**Conclusion and Future Works**

The digitization of East Asian text sometimes involves the creation of unencoded ideographs. In this digital publication age, creating glyphs for unencoded ideographs and embedding them within documents is important, but is not an easy task. GlyphWiki provides a collaborative, sharable and user-friendly method to create a wide variety of glyphs, and adapts them to digital text that is normally dependent on PUA character usage, which is not a searchable, combinable nor future-sustainable method.

The advent of HTML5 (CSS3) technologies and the widespread adoption of OpenType technology may change this situation. With OpenType’s ability to compose multiple glyphs into one, or change a glyph into its variants, one can easily handle unencoded glyphs as an IDS or variants, and CSS3’s the ‘font-feature-settings’ property to realize a fully-functional OpenType font.
GlyphWiki-OpenType collaboration enables one to publish an OpenType font with GSUB features. By using CSS3, one can specify a glyph in a unique way, that can be combined with other texts, and can easily produce a scalable font.

On the other hand, because of the freely-modifiable nature of GlyphWiki, there is always a risk that any glyph can be changed at any time in the future. We had to establish a governance system to maintain the consistency of the names and the fonts.

Acknowledgements

The authors are deeply grateful to Dr. Ken Lunde, for giving insightful comments and suggestions on this paper, and for helping with our attempt to use the various AFDKO tools.

References


ISO/IEC JTC1, ISO/IEC 10646:2012, Information technology -- Universal Coded Character Set (UCS)

